CpG Methylation Imputation

Abhinav Khanna
Student Researcher
akhanna@princeton.edu

Amandeep Singh
Student Researcher
assaini@princeton.edu

Abstract

In this assignment, we try to predict the missing values of an unknown tissue using the values observed for 33 other tissues. We do this by trying a number of different methods, involving weighted means, beta distributions, and various regressors. We evaluated the results of our predictions with the actual observed values at the site of the unknown tissue. We find that taking the weighted mean across the top 5 correlated vectors produced the best result, with the lowest RMSE, 6.3%, as compared to other, more complex methods which resulted in higher error and lower correlation with the actual values.

1 Introduction

The study of the human genome is incredibly immense, and has been making many strides in recent times. Methylation is a very studied topic in genomics. One of the reasons why it is so closely studied is because it provides much insight into epigenetic markers, which are crucial identifiers of potential diseases that could occur in the human body. However, it can be an extremely costly process to accurately record values at every loci in question for every tissue under investigation. Such chips are both monetarily expensive and time wise very expensive. Thus, researchers have been trying to identify ways to use large data sets of prerecorded data to predict the methylation values of a tissue with only 2% of the values recorded. In this paper, we plan to build upon the research that has been conducted and explore ways that we may improve the imputation pipeline to gain better results.

2 Related Work

Currently, there exists a state of the art method of finding methylation values across an entire genome, and that is called whole genome bisulfite sequencing (WGBS). However, WGBS proves to be too costly to realistically be used. Although genome sequencing is making great strides every year, and is even outpacing moores law with methods like Next Generation Sequencing (NGS)[3], it is still costly, and so, using machine learning to predict values is still a necessary tool. Some tools that are coupled with machine learning are the Infinium HumanMethylation 450K BeadChip Kit, which checks roughly 480,000 methylation sites per sample at pre-selected CpG sites. [4]. Much research has been done in the area of using machine learning to predict values. For example, Zhang, Spector, and Engelhardt used random forest classification to achieve 92% prediction accuracy of genome wide methylation levels[1]. This is quite good, and so we hoped to try to achieve a similar accuracy.

3 Methods

3.1 Description of data

The training data used in this paper comes from a set of 33 reference samples that have their entire genome bisulfite sequence. We primarily focus on the data for chromosome 1 for our analysis. The data consists of 33 reference samples with their methlyation proportions for each CpG site on the chromosome. We utilize this information to impute the values for the missing methlyation values in the test chromosome sample.

To better understand the relationships between different tissues, we computed the Pearson R correlation coefficient (Figure 2). Certain tissues have a meaningful, non-pure chance, correlation with other tissues suggesting that certain tissues may be more apt at predicting given test values than other tissues. We also know from Figure 1 that locations that are closer together have higher correlations, and are more similar to each other than locations that are further apart on the chromosome.

Figure 1: Pearson R Correlation for neighboring index locations

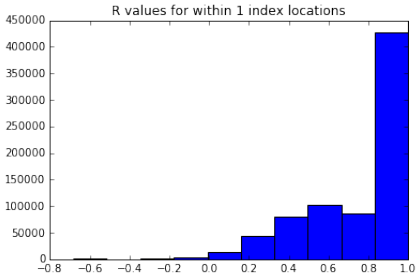


Figure 2: Table of Correlation Values for Different Tissues

	0	1	2	3	4
0	(0, 0)	(0.712834867218, 0.0)	(0.757290076957, 0.0)	(0.708640674253, 0.0)	(0.377274968645, 0.0)
1	(0.712834867218, 0.0)	(0, 0)	(0.703314264614, 0.0)	(0.706602501721, 0.0)	(0.329787127457, 0.0)
2	(0.757290076957, 0.0)	(0.703314264614, 0.0)	(0, 0)	(0.738807452418, 0.0)	(0.351294390129, 0.0)
3	(0.708640674253, 0.0)	(0.706602501721, 0.0)	(0.738807452418, 0.0)	(0, 0)	(0.312640590975, 0.0)
4	(0.377274968645, 0.0)	(0.329787127457, 0.0)	(0.351294390129, 0.0)	(0.312640590975, 0.0)	(0, 0)

Examining the distribution of the mean methylation values across different tissues the distribution appears to have a right skew, and looks like it could be representative of a beta distribution. We believe that attempting to model the similarity between our test tissue and the training tissues as a problem of judging which prior beta distribution the test sample is most likely to have been drawn from may be a worthwhile approach. The mean methylation values can be seen in Figure 3. Furthermore, plotting the methylation values across a given tissue by location yields a graph that does not look linear, suggesting that a non-linear regressor may yield a better result (Figure 4).

3.2 Imputation Pipelines

The basic structure of our approach remained the same across pipelines. We began by calculating the similarity between the test tissue and the training tissues. The similarity was judged through some type of similarity metric, and different pipelines used different metrics. After measuring similarity, the pipelines combined the information from the most similar tissues to predict the values of the test tissue. The process of combining the information was also varied from pipeline to pipeline.

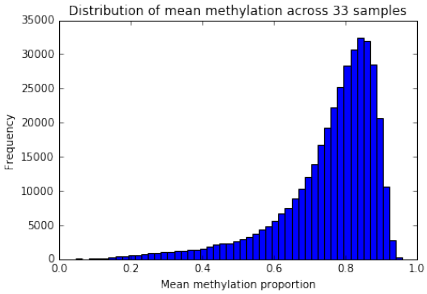


Figure 3: Mean Methylation across 33 Tissues

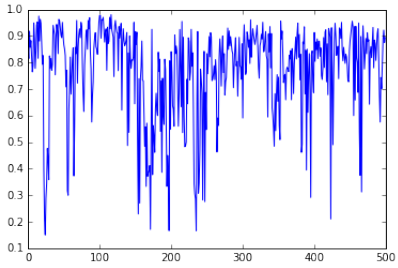


Figure 4: Methylation Values across a Single Tissue

3.2.1 Similarity Metrics Tested

1. Beta Distribution Prior Probability
2. Pearson's R correlation

3.2.2 Combination Methods Tested

1. Weighted mean
2. Support Vector Regression
3. Decision Tree Regression
4. Random Forest Regression
5. Linear Regression

The pipelines that were run with the beta prior segmented the data into window size chunks, and used those chunks to train the regressors for the given location values.

3.3 Evaluation

The effectiveness of each pipeline was judged by examining the Pearson R correlation of the filled in test vector to the full vector, and computing the root mean squared error between the two vectors.

4 Results

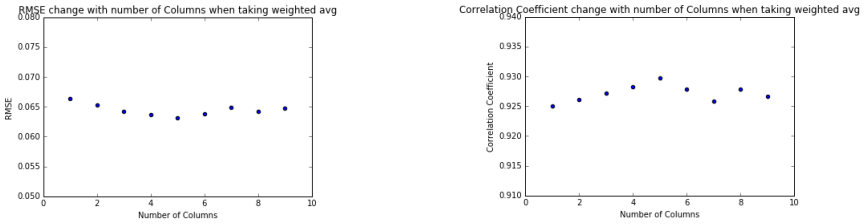
We tested the pipelines against chromosome 1, and graded the accuracy of the imputation based on the RMSE, and Pearson's R correlation values. The following error rates were achieved by predicting the test sample for chromosome 1. Performance was variable with the weighted mean with correlation similarity performing the best out of all the pipelines. In the table, BPS stands for Beta Prior Similarity, SVR stands for Support Vector Regressor, DT stands for Decision Tree Regressor, RF stands for Random Forest Regressor, LR stands for Linear Regression, CS stands for Correlation Similarity, and WM stands for Weighted Mean similarity.

Imputation Accuracy		
Pipeline	**RMSE %**	**Pearson's R**
BPS + SVR	50.0	0.3
BPS + DT	17.7	0.58
BPS + RF	17.7	0.58
CS + WM	6.3	0.93
CS + DT	10.2	0.93
CS + LR	26.2	0.03

For the weighted mean with correlation similarity pipeline, we optimized how many most similar vectors we looked at in order to impute the test value by graphing the change in the RMSE and the Pearson's R correlation with different top most similar vectors. As figure 5 showcases, the lowest

RMSE and the highest correlation occur when 5 columns are considered. For the beta prior similarity pipelines, we segmented the data sets by a given window size, it was predicted that smaller window sizes would perform better because the data would appear more linear on smaller windows, and through trial and error, it was discovered that window size had no effect on the final accuracy.

Figure 5: RMSE and Pearson R change with the number of columns



5 Discussion and Conclusion

Once we observed that certain column vectors had strong correlation, we found the top 3 vectors correlated with each tissue and imputed the NANs by taking a weighted mean across the respective chromosome site. The results we obtained using just the simple method of imputing the values of the NANs by using the values at that location in the most highly correlated tissue, which happened to be tissue 23, was already producing better results than the example code that was provided. And as mentioned above, we found the global minimum of RMSE and the global maximum of Correlation Coefficient when we used the top 5 correlated vectors. And this makes sense, because if we use too few columns to compute the average, any chromosome locations that are outliers in the highly correlated tissue could cause a bad prediction, whereas too many columns when used to take the average can introduce noise and superfluous data, especially since not all tissues correlate well.

Thus, when we used 5 columns to take the weighted mean to impute the missing values in the test set, it produced significantly better results. However, we were determined to do even better, and so instead of taking a simple weighted average, we used methods of regression coupled with weighted averaging to try to predict the values. As seen in the data from the previous section, linear regression performed very poorly, but this was expected, because, as mentioned earlier, the data in each column did not follow a linear trend. Polynomial regression performed better than linear, and decision tree regression even more so, but none performed quite as well as taking simply the weighted average itself across the most correlated columns. And this makes sense: regression in this case puts us at a loss of information, because it tries to generalize a tissue to a trend, but we already have a majority of the data in each tissue, since the NANs are so sparse in the training set. Thus, when we regress and predict a value at a chromosome site, we will incur more error than if we were to simply read the data at the row and use the observed value with some weight. And so, in this example, in terms of complexity providing better results, we notice that less is more: the easier and more straightforward approach yielded better results, and with good reason.

Similarly, when we tried to use the beta distribution as a prior, we discovered that the vectors that were chosen as the most similar are not as close as the vectors that are chosen with the correlation metric. We believe that this may be part of the reason we see far higher error and lower correlation results for all the pipelines based on beta distribution priors. We hypothesize that the beta priors are not dissimilar enough between different tissues to accurately draw the true closest tissues.

All in all, the weighted mean imputation with the correlation measure used for similarity performed the best. If we had more time, we would try to understand how batch size affects similarity. Perhaps, certain areas of chromosomes are better correlated with other areas of different tissues, rather than looking at a chromosome at the entire tissue level. Breaking down the chromosome into smaller parts may yield more refined results even with just weighted average being used.

Acknowledgments

We would like to acknowledge Princeton University for offering the wonderful COS 424 class that provided us with the opportunity to take on this research. We would also like to thank our Professor, Professor Englehart, and her TA team for providing us with guidance and the tools necessary to make this project happen.

References

[1] - Weiwei Zhang, Tim D Spector, Panos Deloukas, Jordana T Bell, and Barbara E Engelhardt. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. arXiv preprint arXiv:1308.2134, 2013

[2] - Peter W Laird. Principles and challenges of genomewide DNA methylation analysis. Nature reviews. Genetics, 11(3):191203, March 2010. ISSN 1471-0064. doi: 10.1038/nrg2732. URL http://www.ncbi.nlm.nih.gov/pubmed/20125086.

[3] - Next-Generation Sequencing (NGS). (n.d.). Retrieved March 22, 2016, from http://www.illumina.com/technology/next-generation-sequencing.html

[4] - Bjo/methylation_imputation. (n.d.). Retrieved March 22, 2016, from https://github.com/bjo/methylation_imputation