COS 424 Homework 2

Vladimir Feinberg
Princeton University
`vyf@princeton.edu`

Abstract

We present an approach for imputing methylation values, a form of epigenetic marker, in tissue samples by relying on methylation values of other tissues and cis-regulatory genetic contexts. A combined collaborative filtering and elastic net linear model approach offer an out-of-sample R^2 of 0.874534 for predicting the methylation values of a tissue sample with only 2% of the values known - there are a total of about 380K sites whose values must be predicted via regression.

1 Introduction and Related Work

Imputation is a general statistical process involving replacing unknown values with a best guess based on a given context. Formally, we consider some vectors $\mathbf{x}_i \in S^n$ for some set S. If we take these $\mathbf{x}_i$ to be sampled from a distribution over S^n, we can make some predictions about the posterior of a $\mathbf{x}$ given some limited observation into a set of $\{\mathbf{x}_i\}$.

In our case, S is a tuple over various biological information: chromosome number, genomic offset, strand type (3'-5' or 5'-3'), and methylation, the proportion of sites in a laboratory sample that had an additional methyl group attached to a cytosine molecule. The methylation count in our data set may be missing information.

Predicting whether the proportion of methylated sites from a sample is at least half has been done with over 90% accuracy by relying on methylation levels of nearby sites, sequence-encoded information such as genetic context, and genomic position to the extent that the site is co-located (in the binary sense) with a DNA sequence for a particular protein or a sequence for a cis-regulartory element (CRE) [11]. Other studies have also found co-localization between CREs and CpGs (sites where the methylation occurs) [12].

2 Exploratory Data Analysis

We focus our analysis mainly on chromosome 1, which has samples from 379551 sites. There are 34 samples for each site, each taken from an expensive WGBS procedure, this method is able to measure about 91% of sites [5]. Our test sample only has about 2% of the sites available from a cheaper procedure: methylation microarrays [11].

We are to use the limited information to impute the missing values in the sample. Because we can only use WGBS as the true value, there are some sites that we don't know the correct value for, even in testing. The imputation procedure may *not* be specified to only impute the sites we are testing for because this is a facet of our testing procedure. The methods constructed attempt to impute the entire genome. We are only able to provide an estimate for our testing error because we can only observe some of the true methylation values with the more expensive technique.

The sites available at test time are consistent - the technology used to assay the methylation values samples sites consistently [3]. This opens up opportunities to take advantage of learning patterns particular to the sites consistently tested by the methylation microarrays - for instance, we can mask one of the well-sampled tissues enabling a sort of "cross-validation" on a reduced feature set.
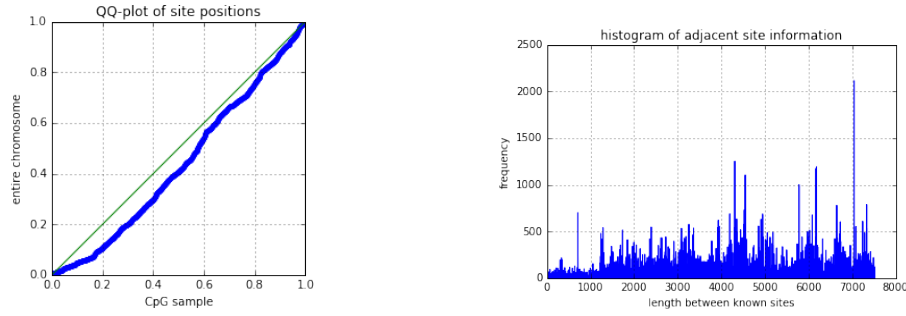
1

Figure 1: The above demonstrates the distribution of the known sites in the test sample. The mean distance between sites is 50.3, with a standard deviation of 87.0. The QQ-plot demonstrates that the sampling is fairly uniform throughout the chromosome.

The microarray may only provide about 2% of the chromosome's information, but it may give enough to find chromosome-wide patterns as in Figure 1. The uniformity here, coupled with the observations of correlation amongst neighboring sites from Figure 2 informs us local values may be informative.

Figure 2: The above demonstrates estimates for the correlation of methylation values with the neighbours a given distance away from a sample site on the same tissue sample. The correlation across 33 of the 34 given tissues was taken. This was performed on 1000 random sites. The last chromosome was dropped due to its 98% sparsity. The "bottoming out" of correlation at about 0.25 as we distance ourselves from the site matches the observed background correlation from [11]. Error bars are $1\hat{\sigma}$.
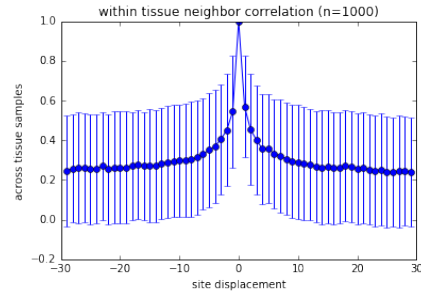
We explore whether there's potential for (1) prediction based on genomic location or (2) prediction based on similarity to other chromosomes. Figure 3 shows the rolling mean of the methylation proportions - it confirms that while there are no obvious trends as a function of sequence, there are chromosomes with similar behavior.
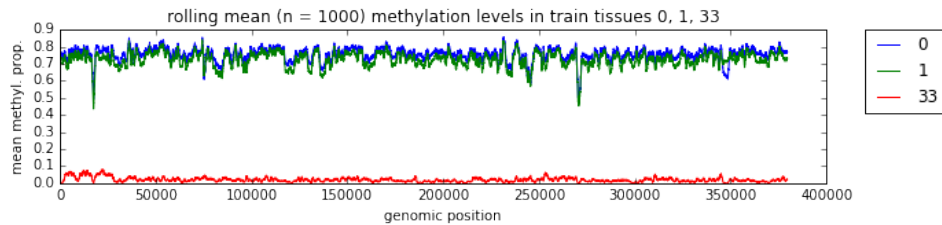

Figure 3: Rolling mean with 1000 bp windows of methylation values for varying tissues. In order, the standard deviation ratio from the original tissue methylation values to the rolled mean ones are 4.83, 4.62, and 11.34. Note that this is the square root of the inverse of proportion of variance maintained by the rolling mean.

As recommended by [11], additional info from the ENCODE project was retrieved, corresponding to the indicators that a given site is within the context of a transcription factor CRE. This introduces 161 binary features, which are used alongside the strand direction information as additional inputs compared to the other tissue methylation samples alone [1].

3 Methods

Recall there are 34 sample tissues. Models were evaluated by average cross-validation performance by supplying 33 tissues with near-full methylation information and 1 sample with only the methylation sites revealed by the 2%-sampling microarray mechanism [3]. However, only 6 folds were performed, because certain tissue samples were sparse or highly uncorrelated with the any of the rest, and would be poor candidates for model selection for our test chromosome, see Figure 4).
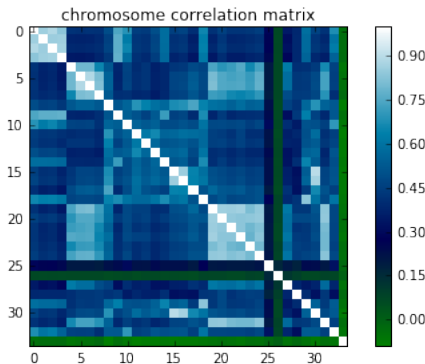
Figure 4: Using only the training data, it was found that the closes training sample was chromosome 19. Thus, we estimate from the diagram above that chromosomes 19-24, inclusive, will be good folds to train our models on (by pretending they each only have limited methylation information).

A fundamental assumption here is that the additional tissue data at test time on the holdout tissue sample will not significantly worsen the top model from cross-validation. This is a fair assumption since the model was selected with 33 similar features.

The observations from the previous sections inform us that several approaches may be successful: interpolation between neighboring sites and identifying clusters of chromosomes according to methylation patterns. The latter has biological basis: the TA mentioned that chromosomes from the same tissue groups have similar methylation. One approach is to use our restricted amount of methylation knowledge to detect what linear combination of tissue samples, when viewed as vectors in the 380K-dimensional site space, best approximates the test sample.

There are many strategies for finding the aforementioned combination. Figure 4 displays several clusters - a Gaussian Mixture Model (GMM) approach could be successful, but was intractable with constructed toolkits given the high dimensionality. A simpler approach that leveraged the cluster structure is the memory-based collaborative filtering approach from Wikipedia, which returns a weighted sum from the fully-assayed samples using a similarity metric defined only on the intersection of the known methylation values [10]. Similarly, we can learn this linear combination by fitting a linear model - in this case, we consider every site index a separate, independent data point, and train on the known methylation values. The difficulty here is that we are using only 2% of the information to learn a model that is then applied to 98% of it. We rely on the fact displayed by Figure 1 - even though this model is learned on a limited space, we are performing interpolation with our regression rather than extrapolation.

Linear models considered were OLS, ridge, lasso, random forest (which is mostly linear), and elastic net. The sparse features, such as the indicators, would be harshly penalized by lasso alone. However, lasso still outperformed ridge due to its restriction of overfitting. As such, elastic net was attempted as a middle ground. Indeed, it outperformed the other models (only its results are displayed in the summary table, Figure 5).

All hyperparameter were trained on a nested 8-fold CV loop, which split the test sample's methylation sites even further for training. The "known tissues" still had some missing samples - these were imputed with the mean value of each individual sample. Fancier imputation was not necessary because there were so few missing sites - interpolation-based imputation yielded the same results.

3

4 Results

method	RMSE	methylation accuracy	R^2
local site interpolation	0.204777	0.889799	-0.613112
within-tissue mean	0.162478	0.937630	-0.015392
across-tissue mean	0.131869	0.953696	0.325614
1 nearest neighbor	0.091203	0.968845	0.677763
cosine mCF	0.130761	0.954496	0.336823
canberra mCF	0.128541	0.955625	0.359184
correlation mCF	0.102694	0.963925	0.591949
tissue elastic net	0.069042	0.976348	0.814531
metadata elastic net	0.154595	0.938566	0.080378
ensemble elastic net	0.068833	0.976385	0.815618

Figure 5: RMSE and R^2 values are calculated on the test data - for this reason R^2 is not adjusted and may also be negative. The methylation accuracy is the performance of the regression when reduced to the 0-1 classification problem for whether a site is methylated. Metadata refers to strand direction and CRE indicators. mCF is the aforementioned memory-based collaborative filtering algorithm, which accepts as input a similarity measure. The ensemble elastic net combines the methylation, metadata, and correlation-based mCF values as explanatory variables.

The three strongest non-ensemble models were combined after observation that their residuals on one of the folds are weakly correlated - the largest R^2 amongst the 3 pairs was 0.641. Taking our best best model, the ensemble elastic net, and making a finer grid search based on the optimal hyperparameters from the CV performed for Figure 5's results. Its results are available in Figures 6 and 7.

Figure 6: Ensemble elastic net regression model residuals on the holdout sample. RMSE is 0.059897, methylation prediction accuracy is 0.980256, and R^2 is 0.874534. Note the residuals are satisfactorily normal.
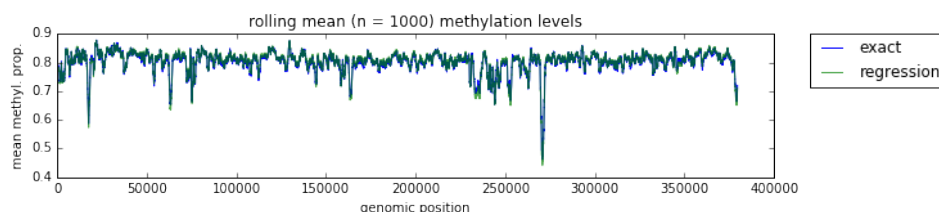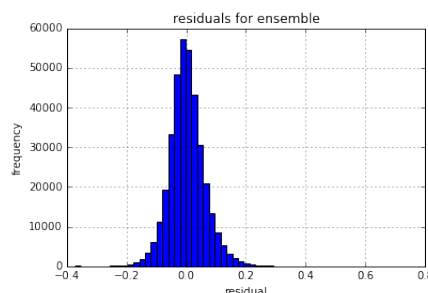

Figure 7: Rolling means of both the final regression model and the exact chromosome, with all indices included. The exact sample is still missing about 1% of sites, which were locally interpolated for this image. The rolling mean computation reduces the variance of both sequences by at most a factor of 22.

5 Conclusion

This report demonstrates the predictive power for methylation values, both individually and when combined, of memory-based collaborative filtering with a correlation-based similarity metric, CRE

context indicators, and methylation values of sites in the same position in the genomic sequence for different tissues. The final R^2 of 0.874 can still be improved.

Colocated site values were not used except in the interpolation imputation. Adjacent methylation values from different tissues could be added as additional regressors.

Furthermore, other methods that were either intractable or unavailable may have promise. For instance, beta regression has a more appropriate range for regression prediction. An iterative, regularized, SVD collaborative filtering algorithm may elicit different patterns than the simple memory-based one [8].

Perhaps the most promising model that encodes all patterns that have been observed, which the author would have needed to implement in TensorFlow (but did not have time for) is a K-cluster GMM over the entire 380K-dimensional methylation vector of each tissue. The variance matrix would agree with our colocated correlation observations - it would be restricted to a diagonal matrix, with values for the $+i$ and $-i$ bands adjacent to the diagonal corresponding to the neighboring sites of displacement i. Bands with $i > A$ would be held as zero. After training for a given K, A, one is equipped with a mixture model over all 380K sites. This allows for a closed-form posterior given only 2% of the sites, the mean of which may be used for prediction (and the variance for confidence).

Acknowledgments

References

[1] CONSORTIUM, E. P., ET AL. The encode (encyclopedia of dna elements) project. *Science 306*, 5696 (2004), 636–640.

[2] HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering 9*, 3 (2007), 90–95.

[3] ILLUMINA. Ifinium HumanMethylation450 BeadChip. Data Sheet, March 2012.

[4] JONES, E., OLIPHANT, T., PETERSON, P., ET AL. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2016-02-23].

[5] LAIRD, P. W. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics 11*, 3 (2010), 191–203.

[6] MCKINNEY, W. pandas: a foundational python library for data analysis and statistics.

[7] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[8] PERCY, M. Collaborative filtering for netflix, Sept. 2009.

[9] PÉREZ, F., AND GRANGER, B. E. IPython: a system for interactive scientific computing. *Computing in Science and Engineering 9*, 3 (May 2007), 21–29.

[10] WIKIPEDIA. Collaborative filtering — Wikipedia, the free encyclopedia, 2016. [Online; accessed 19-March-2016].

[11] ZHANG, W., SPECTOR, T. D., DELOUKAS, P., BELL, J. T., AND THA ONE AND ONLY - THE BEE FROM THE P - DR. BARB ENGELHARDT. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome Biol 16* (2015), 14.

[12] ZILLER, M. J., GU, H., MÜLLER, F., DONAGHEY, J., TSAI, L. T.-Y., KOHLBACHER, O., DE JAGER, P. L., ROSEN, E. D., BENNETT, D. A., BERNSTEIN, B. E., ET AL. Charting a dynamic dna methylation landscape of the human genome. *Nature 500*, 7463 (2013), 477–481.