COS 424 Homework 2

Vladimir Feinberg
Princeton University
vyf@princeton.edu

Abstract

abstract

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) [4]. Other studies have also found co-localization between CREs and CpGs (sites where the methylation occurs) [5].

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 [3]. Our test sample only has about 2% of the sites available from a cheaper procedure: methylation microarrays [4].

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 [2]. 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.

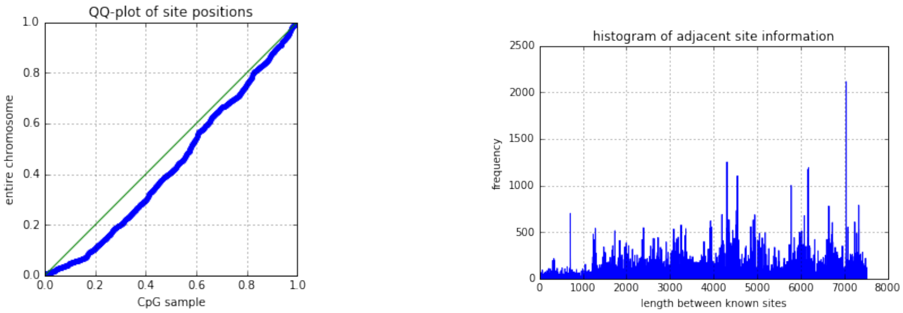

Figure 1: The above demonstrates the distribution of the known sites in the test sample. The mean distance between sites is 50.3, with a standard deviation of 87.0. The QQ-plot demonstrates that the sampling is fairly uniform throughout the chromosome.

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 2. The uniformity here, coupled with the observations of correlation amongst neighboring sites from Figure 2 informs us local values may be informative.

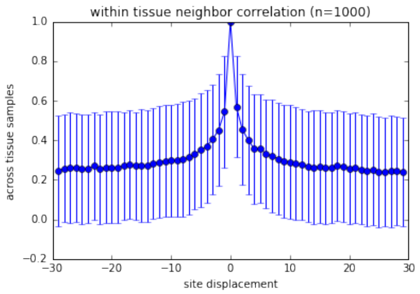

Figure 2: The above demonstrates estimates for the correlation of methylation values with the neighbours a given distance away from a sample site on the same tissue sample. The correlation across 33 of the 34 given tissues was taken. This was performed on 1000 random sites. The last chromosome was dropped 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 [4]. 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 2 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.

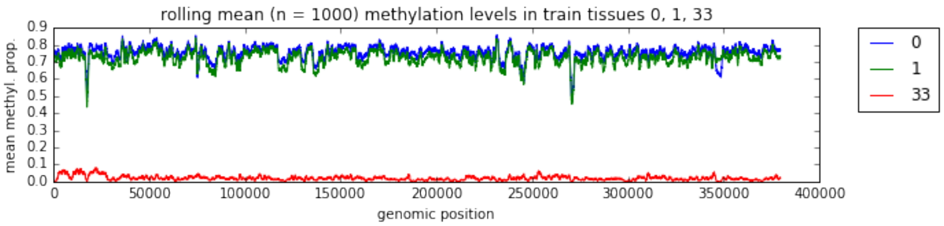

Figure 3: Rolling mean with 1000 bp windows of methylation values for varying tissues. In order, the standard deviation ratio from the original tissue methylation values to the rolled mean ones are 4.83, 4.62, and 11.34. Note that this is the square root of the inverse of proportion of variance maintained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

As recommended by [4], 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 [2]. 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 3).

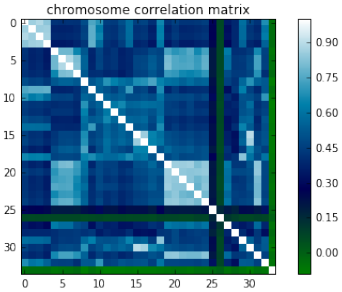

Figure 4: Using only the training data, it was found that the closes training sample was chromosome 19. Thus, we estimate from the diagram above that chromosomes 19-24, inclusive, will be good folds to train our models on (by pretending they each only have limited methylation information).

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 of the promise

4 Results

5 Conclusion

Acknowledgments

The author would like to thank Prof. Engelhardt and Brian Jo for their advice on this project and provision of data. Also, Princeton University and the ENCODE project are responsible access to certain references and additional data sets [1].

References

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

[2] illumina. Ifinium HumanMethylation450 BeadChip. Data Sheet, March 2012.

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

[4] 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.

[5] 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.