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, DNA sequence properties such as ENCODE, 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) [3]. Other studies have also found co-localization between CREs and CpGs (sites where the methylation occurs) [4].

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

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 [1]. 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.

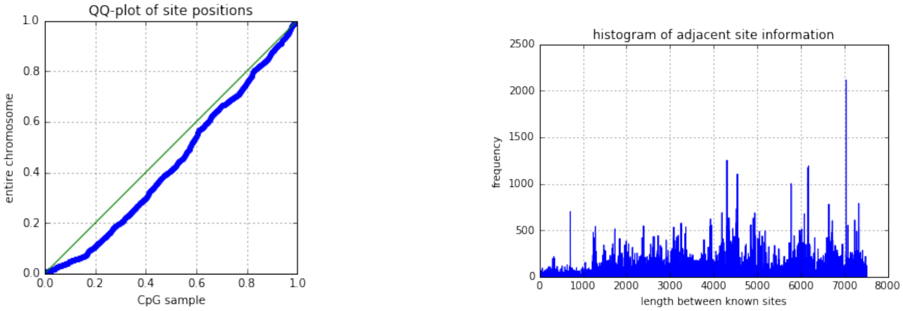

Figure 1: The above demonstrates the distribution of the known sites in the test sample. The mean distance between sites is 50.3, with a standard deviation of 87.0. The QQ-plot demonstrates that the sampling is fairly uniform throughout the chromosome.

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.

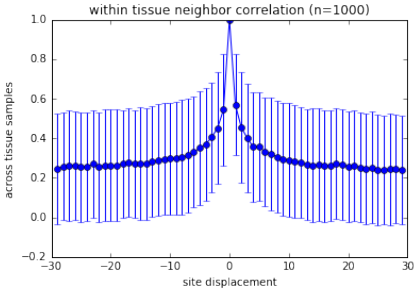

Figure 2: The above demonstrates estimates for the correlation of methylation values with the neighbours a given distance away from a sample site on the same tissue sample. The correlation across 33 of the 34 given tissues was taken. This was performed on 1000 random sites. The last chromosome was dropped 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 [3]. 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.

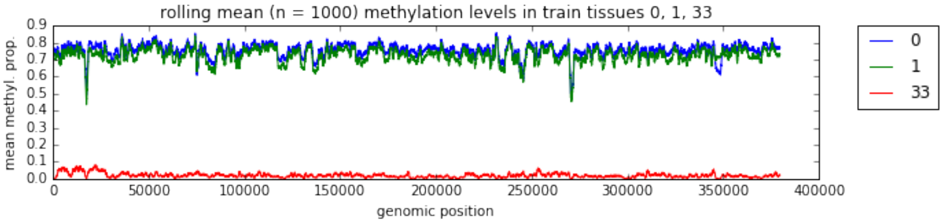

Figure 3: Rolling mean with 1000 bp windows of methylation values for varying tissues.

Finally, we discuss the sparsity of the chromosomes - all 34 training tissues except the last have no methylation proportions (betas) under 0.01 for any site. The last has about 98% such values.

Of the limited amount of known sites for the test chromosome, none of our sites have such low methylations either. In the following, the last chromosome will be excluded from certain aggregates to avoid skewing certain statistics. However, it may still prove useful for prediction, and may be included as an explanatory variable.

3 Methods

4 Results

5 Discussion and Conclusion

Acknowledgments

References

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

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

[3] 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.

[4] 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.