

ASSIGNMENT 2: IMPUTING METHYLATION STATUS

Barbara Engelhardt, Princeton University

out 03/10/2015; due 04/02/2015

Background

Much of the variance, or differences, in traits across individuals is regulated by genetic and epigenetic markers. While genetics is the study of DNA, which is inherited, stable through cell division, and (generally) does not change in response to environment, *epigenetics* is the study of non-genetic cellular processes that may be inherited, are stable through cell division, and may change in response to external and internal cellular stimuli. Epigenetic markers may change within an individual over time and may be variable across different tissues in the body. DNA methylation is probably the best studied epigenetic modification of DNA, but our understanding of DNA methylation is poorly understood. In vertebrates, DNA methylation occurs by adding a methyl group to the fifth carbon of the cytosine residue, mainly in the context of neighboring cytosine (C) and guanine (G) nucleotides in the genome (*CpG sites*). DNA methylation has been shown to play an important functional role in the cell, including involvement in DNA replication and gene transcription, with substantial downstream association with development, aging, and cancer.

The important role of methylation in cellular processes imply that characterizing genome-wide DNA methylation patterns is necessary to better understand the regulatory mechanisms of this epigenetic phenomenon Laird [2010]. Recent advances in methylation-specific microarray and sequencing technologies have enabled the assay of DNA methylation patterns genome-wide at single base-pair resolution Laird [2010]. The current gold standard to quantify single site DNA methylation levels across a genome is whole genome bisulfite sequencing (WGBS), which quantifies DNA methylation levels at ~ 26 million (out of 28 million total) CpG sites in the human genome. However, WGBS is prohibitively expensive for most current studies Laird [2010]. As an alternative, methylation microarrays, and the Illumina HumanMethylation 450K Beadchip in particular, measure bisulphite treated DNA methylation levels at $\sim 482,000$ preselected CpG sites genome-wide Zhang et al. [2013]; however, these arrays assay less than 2% of CpG sites, and this percentage is biased to gene regions and CGIs. The purpose of this project is to develop quantitative methods to predict (*impute*) methylation status at unassayed CpG sites across the genome.

Project definition

Your goal in this homework project is to use a data set consisting of 37 reference samples with whole genome bisulfite sequence (WGBS) data Ziller et al. [2013] to impute (predict) the methylation levels for a sample that only has a small subset of observed methylation levels (i.e.,

only those CpG sites on the Illumina 450K array). We have supplied you the reference data separated out by chromosome. Please feel free to **only use the data from chromosome 1 for this project**. Here are the steps to get started:

- Get access to the Nobel Cluster (automated):
 - Please submit the registration form here:
<https://www.princeton.edu/researchcomputing/computational-hardware/nobel/registration/>
 - Make sure to first enable your OIT Linux account by following the link on that page.
 - If you do not have access to this cluster, please email cses@princeton.edu and let us know.
- Log on using your NetID to `nobel.princeton.edu`;
- We have created a read-only directory `/u/biancad/COS424` that contains all of the data files for this project. Please feel free to leave the data here and write scripts that access these data.
- In this directory there is a README file that describes the format of each of the files.
- Train your imputation method using both the `*train*` and the `*sample*` files; evaluate your predictions using the `*test*` file. Feel free to restrict to `*chr1_*` files only. Clearly document what files you are training on.
- Note that there are nans (missing data) in these files. You are responsible for predicting every column of the `sample` file with the final column value = 0. Some of the training data and some of the test data will be missing values. Be sure to take this into account in your data analysis.

In the train file, you will find position numbers corresponding to the genomic locations of each of the CpG sites. There will be 37 values that correspond to each of the 37 reference samples; these values $\beta \in [0, 1]$ are the fraction of reads from this sample that are methylated. In general, most CpG sites have values of β that are greater than 0.7, but in some regions of the genome this will not be the case. In this project, you will use the information contained in the reference (train) file to *impute*, or predict, the β values for every CpG site (position) that is in the sample file but has a current value of nan and has a final column value of 0. Because we have the 'ground truth' for the test sample in the test file, with β values for every CpG site, you can compare the quality of your predictions against the known β values. Do not use these ground truth values in the imputation step, as the purpose of this exercise is to predict these values from a subset of the data and the reference sequences.

Then, you should build (multiple) methods for predicting the methylation levels at each CpG site that fits a model to the reference data set. First, you should carefully design the set of *features* that you will use to predict methylation values at each position. These should include, for example, the β values at that position for the 37 reference samples; you might also include the values of the nearby positions in the sample file, and perhaps the distance to those neighboring

positions in bases. For the methods, feel free to use the generalized linear regression models we have discussed in class as well as others mentioned in our text books, described in the scientific literature, or implemented in software. You may also use more sophisticated classifiers (see *Extensions*). You might consider using some type of regularization or sample weighting. Finally, you should evaluate the classifiers you apply to this problem according to (at a minimum) r^2 and root mean squared error (RMSE) to compare the imputed levels against the true methylation levels.

We have run the prediction for the samples using *mean value imputation*, where the approximated β value is the mean of the β values in the 37 reference samples at that position. We found that this approach gives us $r^2 = XXX$ and $RMSE = XXX$.

Essential to any data analysis task is the interpretation of the results. What features were most important for prediction, and what do these features tell us about the problem? Were some reference samples more predictive of the held out sample than others? What did the distribution of prediction errors look like: were they approximately Gaussian and zero centered, or do they have a heavy tail, or something else? Simply building a machine learning approach to solve the problem does not constitute a data analysis; recovering and characterizing signal from these results does.

Deliverables

Your deliverables for this project include:

- A four page (not including citations) summary of the project work, which should contain (as described in the Example project write up on Piazza):
 - A title, authors' names, and abstract for the project;
 - an introduction to the problem being addressed;
 - a description of the data;
 - a description of the methods developed and used, and how they were fitted using reference data;
 - a presentation of the results of the methods applied to the test data;
 - a discussion of the results, including specific examples of single CpG sites or features that highlight the behavior of the imputation models;
 - a short summary and conclusion, including extensions that you believe would be particularly valuable based on the results;
 - a *complete* bibliography to support the methylation databases, feature selection, prediction methods, code bases, and related work that are relevant to your project.
- **please make your analysis code available to us: put a single file in the same dropbox folder with the same name as your PDF project writeup (different extension).**

Please use the \LaTeX template we have provided for you. Put your PDF write up of the project into https://dropbox.cs.princeton.edu/COS424_S2015/Assignment2_Methylation_Imputation by 5pm on the assignment due date, with the file name `<author1PUIID>_<author2PUIID>_hw2.pdf`. Please only submit one PDF per pair of authors.

We strongly recommend *writing as you go* in the project, which means starting to write the project report as you are downloading and analyzing the data. That said, you should avoid speculative writing, and only write results once you have them.

Extensions

If you would like to extend this assignment to more interesting ground after first completing the basic deliverables for the project, you might consider the following:

- *Extend the data set*: The data here represents a single chromosome; there are 21 additional chromosomes that you may wish to work with. We have uploaded the full sets of data if you would like to try to take advantage of them. Moreover, there are a number of publicly available WGBS data sets in public data bases; in particular, Gene Expression Omnibus appears to have 51 entries corresponding to public human (*Homo sapiens*) WGBS data as of this writing. Processing and incorporating these data sets as additional reference data—including ones you personally compile—and releasing these data with appropriate permissions would be worthwhile. We are happy to provide you with tools for processing these data; please ask.
- *More interesting features*: while we have only asked you to predict methylation levels using a set of reference sequences, there are many extensions to this to consider, including features involving:
 - *neighboring CpG sites*: there is often correlation in methylation levels among neighboring CpG sites
 - *genetic context*: methylation levels are affected by whether the CpG site is in a CGI, intronic, exonic region
 - *cis-regulatory elements*: methylation levels are correlated with some regulatory elements Zhang et al. [2013]
 - *genomic sequence properties*: methylation levels are correlated with some properties of DNA sequence, including proportions of C and G nucleotides, sequence conservation, and signatures of selection
- *More complex classifiers*: there are a number of more sophisticated statistical models that might be used for this task, including, e.g., Gaussian processes [Roberts et al., 2013, Rasmussen, 2006], regularized beta generalized linear models, or something of your own design that might identify latent structure in the data that is predictive of methylation levels at specific CpG sites.

- *Confidence intervals*: Imputation may be much easier for some CpG sites than for others; quantifying uncertainty in prediction values is an interesting extension that would be useful for downstream tasks.
- *Better evaluation metrics*: the downstream use of imputed methylation levels is to test for association with traits, such as cancer status or type II diabetes. What are the best ways to evaluate your predictions with this downstream task in mind? Can you improve model evaluation using cross validation instead of our training and test sets?
- *Additional types of problems*: Can you use these data to predict when two CpG positions will have different values, or test for differences in methylation status?

Resources

There is a large literature on epigenetic markers. Many current methods for imputing methylation levels at specific CpG sites involve fairly simple classification methods and large numbers of features or reference data sets. Review some of this literature to get ideas on ways to create a really great imputation method.

References

- Peter W Laird. Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics*, 11(3):191–203, March 2010. ISSN 1471-0064. doi: 10.1038/nrg2732. URL <http://www.ncbi.nlm.nih.gov/pubmed/20125086>.
- Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.
- S Roberts, M Osborne, M Ebdon, S Reece, N Gibson, and S Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.
- 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.
- Michael J. Ziller, Hongchang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David a. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, pages 1–5, August 2013. ISSN 0028-0836. doi: 10.1038/nature12433. URL <http://www.nature.com/doifinder/10.1038/nature12433>.