Whole-Genome Bisulfite Sequencing

Lee Carlin

November 18, 2017

1 Framing the Problem

The purpose of this document is to keep track of work and progress, somewhat randomly and unstructured. We are interested in understanding the genome sequencing signals (e.g. methylation measurements, WGBS,...) in terms of variation, noise, genomic regions, inference, etc..

- Signal Smoothing: Local Likelihood, other smoothing functions
- DMRs as a function of different smoothing parameters (multi-parameter/ multi-scalre representation)
- Inference
- Methylation Variation

2 Signal Smoothing

Signal smoothing is often applied to the noisy genome measurements data (WGBS) in order to cope with the measurement noise [1],[2],[4]. Here the goal is to smooth (e.g. local likelihood averaging) the data so inference will be more robust and less sensitive to noise thus improving statistical analysis. In addition, WGBS is costly, and by using smoothing we are able to reduce the amount of coverage otherwise required. Also, different genomic regions exhibit different levels of DNA methylation variation among individuals (biological difference). This in itself can cause us to identify false DMR when not enough replicates are provided. Both costs and biological differences are the driving factor behind smoothing the signal.

The main area of research here is trying to find a more natural and data dependent and flexible smoothing function.

2.1 Local Likelihood Smoother

3 Inference

Summary of [2]

- We assume π_j is characterized by a smooth varying function f that depends on the genomic location: $\pi_j = f(l_j)$ for location l.

- We smooth each sample using a local-likelihood smoother, using a window size $h(l_j)$.
- The window size must be at least 2kb wide and include 70 CpGs.
- we assume $log[f(lj)/1 - f(lj)]$ is approximated by a second degree polynomial within each window.
- We assume that data follow a binomial distribution and the parameters defining the polynomial are estimated by fitting a weighted generalized linear model to the data inside the genomic window. For data points inside this window, indexed by l_k, weights are inversely proportional to the standard errors of the CpG-level measurements, $\sqrt{\pi_k(1-\pi_k)/N_k}$, and decrease with the distance between the loci $|l_k = l_j|$ according to a tricube kernel

Identification of differentially methylated regions
We compare different groups of samples, and find region showing consistent difference in methylation while taking biological variation into account. In order to that we compute a signal to noise statistic:

- Let $X_i \in 0, 1$ represents the sample from individual i in control or case.
- We assume that the samples are biological replicates within a group.
- denote $M_{i,j}, U_{i,j}, \pi_{i,j} = f_i(l_j)$ the statistics of interest per sample per location.
- we assume $f_i(l_j) = \alpha(l_j) + \beta(l_j)X_i + \epsilon_{i,j}$
- Here $\alpha(l_j)$ represents the baseline methylation profile and $\beta(l_j)$ the true difference between the two groups and our function of interest, with non-zero values associated with DMRs.
- $\epsilon_{i,j}$ represent biological variability with the location-dependent variance $var(\epsilon_{i,j}) = \sigma_j^2$ assumed to be a smooth function.
- Note that increasing coverage does not reduce the variability introduced by Îţ; for this we need to increase the number of biological replicates.
- the smooths profiles estimate the $f_i(\hat{l}_j)5$
- To estimate the smooth location-dependent standard deviation, we first compute the empirical standard deviation across the two groups. we floored these standard deviations at their 75th percentile. To further improve precision, we smoothed the resulting floored values using a running mean with a window size of 101. We get $\sigma(\hat{l}_j)$
- The signal to noise statistics is therefore: $t(l_j) = \frac{\beta(\hat{l}_j)}{\sigma(\hat{l}_j)\sqrt{1/n_1+1/n_2}}$

- DMRs (regions for which $\beta(l_j) \neq 0$) we look for consecutive CpGs such that for all $t(l_j) > c$ or $t(l_j) < -c, c > 0$ for some c.
- CpGs further than 300 bp apart were not permitted to be in the same DMR.
- We recommend including in the procedure only CpGs that have some coverage in most or all samples.
- Furthermore, we recommend filtering the set of DMRs by requiring each DMR to contain at least three CpGs, have an average β of 0.1 or greater, and have at least one CpG every 300 bp.

4 Methylation Variation

Here the objective is to improve the estimation of the methylation variability within/ between samples/groups and/or predict variability of regions using other genomic regions that exhibit similar characteristics.

For example, given covariates $[(\mu_{x_1}, \mu_{x_2}), (d(x_1, x_2)), (\sigma_{x_1}, \sigma_{x_2})), \dots]$ for sample/group x_1, x_2 at different genomic windows, choose a specific window and identify "similar" windows in terms of covariates that will aid in predicting the variability in the chosen window. The repetitive nature of the genome makes this approach somewhat enticing.

The variation problem also relates to the smoothing and genomic regions problems: better estimation of the variation will hopefully induce better inference. Here we use the following studies: [3], [7],

5 Multi-scale Representation

This idea originated from the work done by [5] and [6]. The goal here is to detect genomic regions in different scales (genomic information can be represented on different scales, ranging from bases to thousands of bases).

One research direction that we have is applying Persistent Homology (PH) to find regions at different resolutions by computing the topological features of the information space. PH is usually used to detect somewhat complex topological features that are non-functional in nature such as circles, cones, etc., but in our case we are looking at semi-defined features such as trends and regions represented by the genomic signals. One possible way to modify the approach is to add artificial topology to the information space in order to aid in detecting the features we are looking for by filling in missing but non-interfering structures.

References

[1] Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, page 201602413, 2016.

[2] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83, 2012.

[3] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775, 2011.

[4] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209, 2012.

[5] Theo A Knijnenburg, Stephen A Ramsey, Benjamin P Berman, Kathleen A Kennedy, Arian FA Smit, Lodewyk FA Wessels, Peter W Laird, Alan Aderem, and Ilya Shmulevich. Multiscale representation of genomic signals. *Nature methods*, 11(6):689–694, 2014.

[6] Svetlana Lockwood and Bala Krishnamoorthy. Topological features in cancer gene expression data. *arXiv preprint arXiv:1410.3198*, 2014.

[7] David Siegmund, Benjamin Yakir, and Nancy R Zhang. Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*, pages 645–668, 2011.