B. Tech. Project Report *Phase* I

Expectation Maximization algorithm and Gaussian Mixture Model for analysis of DNA methylation

Submitted in partial fulfillment of requirements
for the award of the degree of Bachelor of Technology from IIT Guwahati

Under the supervision of

Dr. Anil Mukund Limaye

Submitted by

Siddhesh Jitendra Metkar
170106057

November 23, 2020
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, INDIA

Certificate

This is to certify that the work presented in the report entitled **“Expectation Maximization algorithm and Gaussian Mixture Model for analysis of DNA methylation”** by **Siddhesh Jitendra Metkar (170106057),** represents an original work under the guidance of **Dr. Anil Mukund Limaye.** This study has not been submitted elsewhere for a degree.

Signature of student:

Date: November 23, 2020
Place: IIT Guwahati

Siddhesh Jitendra Metkar (170106057)

Signature of supervisor

Date: November 19, 2020
Place: IIT Guwahati

Dr. Anil Mukund Limaye

Signature of HOD

Date: November 19, 2020
Place: IIT Guwahati

Head
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati, India

Contents

1. Abstract

We know that there is an urgent need to identify biomarkers for progressive complex diseases like cancer. The identification of a consistent biomarker from comprehensive genome data is a huge problem. The use of straight statistical measures can result in large false rates, especially in the context of quantitative data like DNA methylation. Hence there is a demand for statistical approaches whose goal is to extract meaningful features by taking out false positives. One such widely used methods in DNA analysis is filtering of data based on variance. The use of graphs (skewness, kurtosis) simplifies the study. The features representing less variable Gaussian distributions are of more interest. This thought of performing feature selection of molecular profiles using Gaussian Mixture Modelling has been discussed in this paper.

A usual degree of DNA methylation is fraction methylation, a measure bounded between 0 and 1, with variance as a feature. These distributional properties have prompted the development of novel statistical strategies for the research of biological hypothesis. The questions are similar, testing for differential DNA methylation between training, or looking for novel subgroups in statistics. Below, we have implemented a statistical strategy that use percentage methylation as the outcome variable for the analysis.

2. Objective

The objectives of this project are: (1) to expand the novel models for the distribution of methylation proportions to select differentially methylated loci between cancers and regular subjects; (2) to advocate a classification method that differentiates tumor subtypes using DNA methylation profiles; (3) to increase laptop software programs or algorithms that implement techniques developed in particular ambitions 1-2.

We hope that this project will enhance current and future efforts in understanding the importance of DNA methylation profiles in cancer and other diseases.

3. Introduction

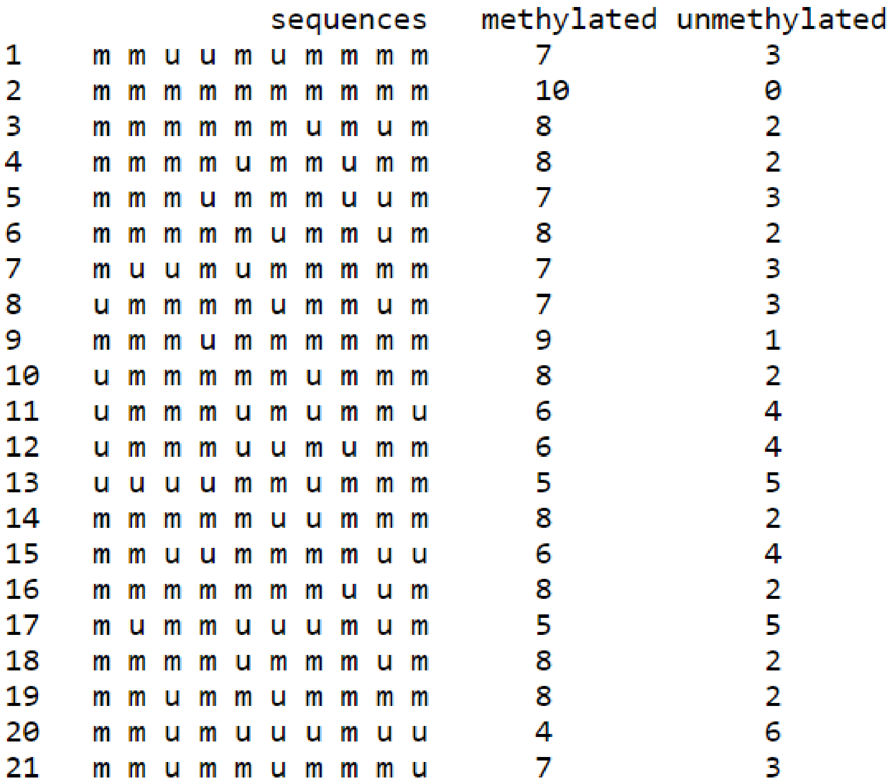
DNA methylation is one of the most widely studied epigenetic changes in mammals. In regular cells, it assures the right regulation of gene expression and stable gene silencing. DNA methylation is related to histone changes and the interaction of these epigenetic adjustments is important to modify the functioning of the genome by means of converting chromatin structure. The covalent addition of a methyl group occurs usually in cytosine within CpG dinucleotides which might be concentrated in huge clusters referred to as CpG islands. DNA methyltransferases are chargeable for organizing and upkeep of methylation pattern. It's typically regarded that inactivation of certain tumor-suppressor genes takes place as a consequence of hypermethylation inside the promoter regions and a severe research have verified a vast range of genes silenced by means of DNA methylation in one of a kind of cancer types. On the other hand, worldwide hypomethylation, inducing genomic instability, also contributes to cell transformation. Aside from DNA methylation alterations in promoter areas and repetitive DNA sequences, this phenomenon is associated additionally with law of expression of noncoding RNAs such as microRNAs which could play role in tumor suppression. DNA methylation appears to be promising in putative translational use in patients and hypermethylated promoters may also function as biomarkers. Furthermore, in contrast to genetic alterations, DNA methylation is reversible what makes it extraordinarily exciting for remedy tactics. The importance of DNA methylation alterations in tumorigenesis encourages us to decode the human epigenome.

The primary purpose of this project is to discover the analogous problem of feature choice within the context of DNA methylation information. Certainly, DNA methylation markers have been proposed as early detection, diagnostic and prognostic markers in a huge range of different diseases which includes cancer. Catalyzing this accelerated interest in epigenomics and enormous advances in technology that now permits the measurement of DNA methylation at over heaps of CpG dinucleotides. We have quantified DNA methylation as u (unmethylated) and m (fully methylated). Despite the fact that normalization and clustering techniques designed for beta-valued DNA methylation have lately been investigated, there's still a scarcity of feature choice techniques.

4. Methods and Materials

4.1 Generation of N number of sequences each having 10 CpG sites with probability that a site is methylated as ‘p’

m = a site is methylated **u** = a site is not methylated



```
                  sequences   methylated unmethylated
1     m m u u m u m m m m       7            3
2     m m m m m m m m m m       10           0
3     m m m m m m u m u m       8            2
4     m m m m u m m u m m       8            2
5     m m m u m m m u u m       7            3
6     m m m m m u m m u m       8            2
7     m u u m u m m m m m       7            3
8     u m m m m u m m u m       7            3
9     m m m u m m m m m m       9            1
10    u m m m m m u m m m       8            2
11    u m m m u m u m m u       6            4
12    u m m m u u m u m m       6            4
13    u u u u m m u m m m       5            5
14    m m m m m u u m m m       8            2
15    m m u u m m m m u u       6            4
16    m m m m m m m u u m       8            2
17    m u m m u u u m u m       5            5
18    m m m m u m m m u m       8            2
19    m m u m m u m m m m       8            2
20    m m u m u u u m u u       4            6
21    m m u m m u m m m u       7            3
```

Fig. 4.1 sequences generated for methylation probability p = 0.7

4.2 Maximum Likelihood Estimation

In statistics, maximum probability estimation (MLE) is a technique of estimating the parameters of a possibility distribution by maximizing a probability function, so that under the assumed statistical version the found information is maximumly probable.

Equation of single variable linear regression is $\mathbf{Y = B_0 + B_1X}$.

Generally, we use Least Square Fit method to estimate parameters B_0 and B_1. But MLE is a method used to estimate the parameters based on our observations.

The Bayes Theorem is given as:

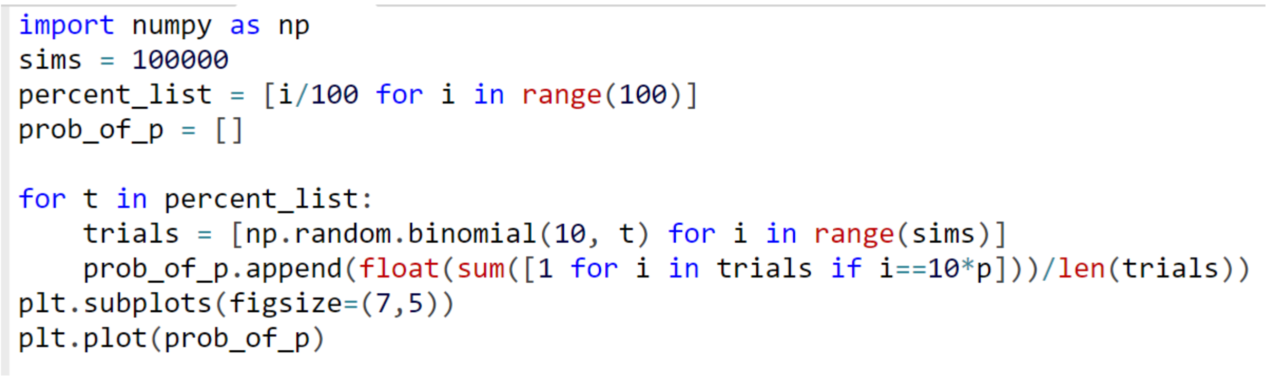
$$\mathrm{P}(\theta \mid x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n \mid \theta)\, \mathrm{P}(\theta)}{\mathrm{P}(x_1, x_2, \ldots, x_n)}$$

Where, P(θ) = distribution of θ parameter

P(x_1, x_2, ………, x_n) = average probability data of all parameters

Hence, to maximize P(θ | x_1, x_2, ………, x_n) we have to maximize f(x_1, x_2, ………, x_n | θ) if we assume P(θ) as uniform distribution.

We did the same using binomial function in Python on our generated data.



```
import numpy as np
sims = 100000
percent_list = [i/100 for i in range(100)]
prob_of_p = []

for t in percent_list:
    trials = [np.random.binomial(10, t) for i in range(sims)]
    prob_of_p.append(float(sum([1 for i in trials if i==10*p]))/len(trials))
plt.subplots(figsize=(7,5))
plt.plot(prob_of_p)
```

The output for probability p = 0.7 is a graph with peak at 70% methylation which was expected

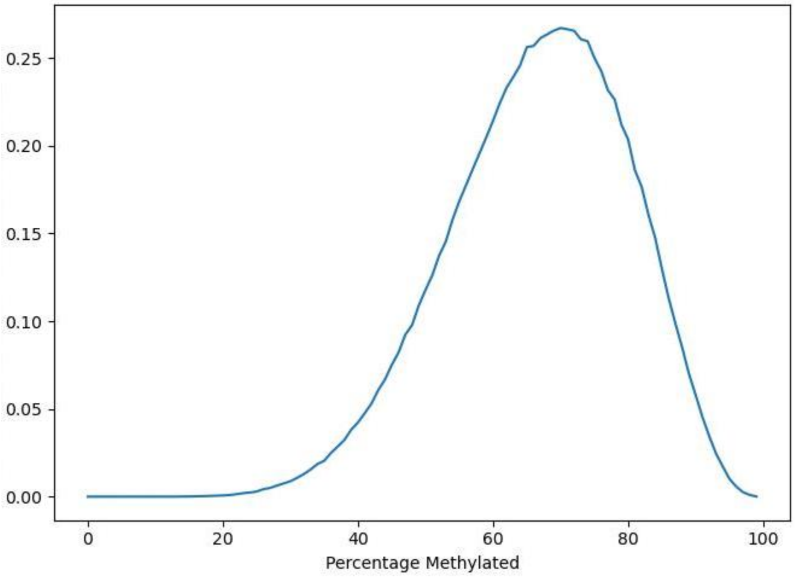


Fig. 4.2 MLE for methylation probability p = 0.7

4.3 Mixture Modelling and Expectation Maximization Algorithm

A problem with MLE is, it is based on assumption that the dataset is complete, this doesn't imply that model has access to all the data. It assumes that all the variables present are related to the problem. But not always the whole dataset is relevant to the problem. There are datasets with some relevant variables, and some non-relevant variables that cannot be observed but are present in the dataset and affect the random variables. These unobserved variables are known as latent variables. The standard MLE will not work if latent variables are present. So, to search for desired parameters Expectation-Maximization approach is used.

Expectation-Maximization or EM algorithm is an iterative algorithm that cycles between two modes namely E-step and M-step. The estimation(E)-step creates a log-likelihood expectation function using current estimated parameters. The maximization(M)-step maximizes the likelihood based on parameters in E-step. We implemented the algorithm on a "normal two-component mixture model".

Mixture of 2 distinct probabilities

We first generated N number of sequences of two different probabilities and shuffled the sequences randomly and plot a histogram of the mixture.

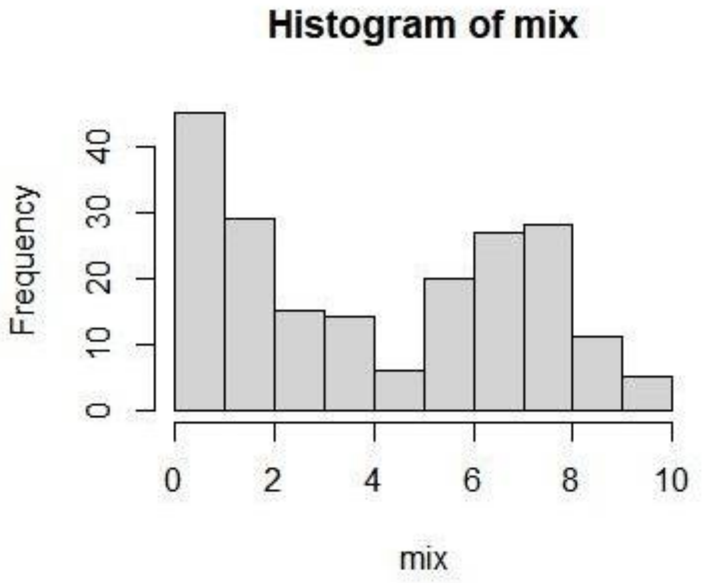


Fig. 4.3 Plot for mixture of sequences with probabilities 0.2 and 0.7

EM algorithm implementation

We begin by choosing some initial parameters. Then in the E-step we compute the probability of each CpG site to be methylated or unmethylated based on the current model parameters.

$$T_{j,i}^{(t)} := \mathrm{P}(Z_i = j \mid X_i = \mathbf{x}_i; \theta^{(t)}) = \frac{\tau_j^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}{\tau_1^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_1^{(t)}, \Sigma_1^{(t)}) + \tau_2^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_2^{(t)}, \Sigma_2^{(t)})}$$

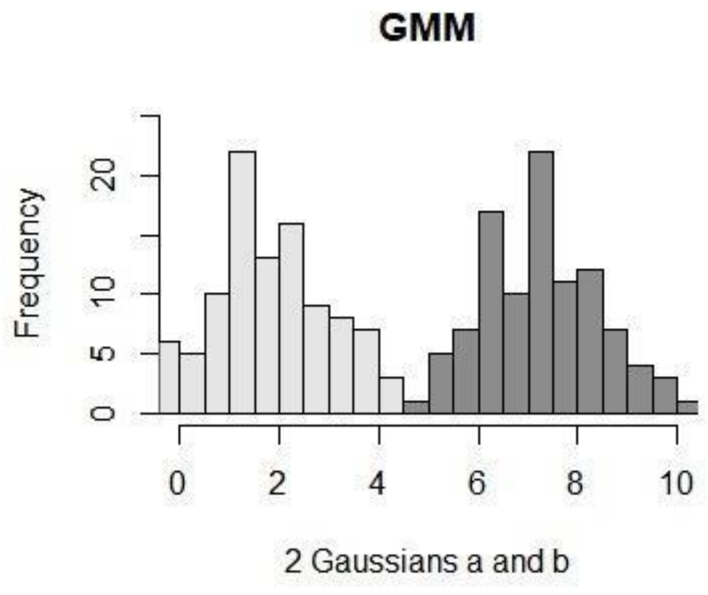
In the M-step, we update the parameters and group proportions by considering the probabilities from the E-step as weights. We compute the Mean and Variance.

$$\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^n T_{1,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n T_{1,i}^{(t)}} \qquad \Sigma_1^{(t+1)} = \frac{\sum_{i=1}^n T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^\top}{\sum_{i=1}^n T_{1,i}^{(t)}}$$

And similarly

$$\boldsymbol{\mu}_2^{(t+1)} = \frac{\sum_{i=1}^n T_{2,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n T_{2,i}^{(t)}} \qquad \Sigma_2^{(t+1)} = \frac{\sum_{i=1}^n T_{2,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})^\top}{\sum_{i=1}^n T_{2,i}^{(t)}}$$

This algorithm was then implemented on the mixture of methylation probabilities 0.2 and 0.7 and the probability distribution of the obtained mean and variance for two gaussians was generated.



5. Results

The EM algorithm was performed on shuffled sequences generated with probabilities 0.2 and 0.7. When 100 sequences for each probability were mixed with initial mean values set as 0.1 and 0.9 for the two gaussians, the output was generated in 59 iterations. The means of the two gaussians in the output were 2.041618 and 7.241509 which is very close to the desired output of 2 and 7 respectively. When the number of sequences was increased (to 1000), more accurate results were obtained, also the number of iterations increased (to 79). It was also observed that when the initial values are taken close to desired probabilities, the number of iterations reduced and accuracy was increased.

Even though Maximum Likelihood Estimation (MLE) and EM can both find “quality-match” parameters, both work in a different fashion. MLE accumulates all the information first after which it uses that information to construct the maximum probably model. Whereas, EM takes a guess on the parameters first accounting for the missing facts, then tweaks the model to improve the guesses and the discovered information.

The primary benefits of the EM algorithm are its simplicity and simplicity of implementation. Unlike, methods like the Newton-Raphson, implementing EM algorithm does not usually require heavy preparatory analytical work. It is straightforward to the software. It reduces to very simple re-estimation formulae and it's miles feasible to apply preferred code to carry out the E-step. The EM algorithm improves a parameter’s estimation via the multi-step system. However, it on occasions desire a few random starts to locate the satisfactory parameters because the algorithm can hone in on a local maximum that isn’t that near the global maximum. That is, it could perform better if you take a right ‘initial guess’ in the first step or iteration.

The EM algorithm also has some limitations. It can be very slow at times. It gives satisfactory results only when some percentage data is latent or missing and the attributes (dimensionality) of the data is not huge. If the dimensionality is increased, the E-step takes time. The global maximization of the auxiliary function finished throughout the M-step can be deceptive. Except a few specific instances, the EM algorithm is not assured to converge to a Global maximizer of the probability.

6. Conclusion and Discussion

We described the method which allows the integrative evaluation of CpG sites with experimental parameters and datasets. Gaussian mixture models is a technique for density estimation wherein the parameters of the distributions are fit using the EM algorithm. The algorithm provides accurate estimates of CpG site modifications even in some confounding datasets. It's also important to additionally factor out that during this work we have most effective explored the EM as a characteristic choice step, i.e. inferring structure in a single-dimensional DNA methylation profiles to pick out true positives more reliably. One can also wish to apply the mixture model to cluster samples over more than one function/dimension. But, analytically, it isn't yet feasible to absolutely comprise the covariance structure of the capabilities inside the inference procedure, which consequently precludes software to cluster more than one dimension. Thus, we always have scope for future investigation.

For the reason that DNA methylation biomarkers for improved diagnosis and/or early detection of cancers are probable to be characterized by means of small impact sizes, it's far crucial to have powerful statistical algorithms in the vicinity that can assist discern true from false positives. Hence, the Gaussian mixture model and Expectation Maximization algorithm discussed above, will be helpful for any study on DNA methylation profiling.

7. Acknowledgments

I would like to thank Dr. Anil Mukund Limaye for being my supervisor, providing support and guidance and clearing all the doubts throughout the project, despite the pandemic situation. I would also like to thank my family members who were always there for support during the completion of this project.

References

1. Interplay of ERα binding and DNA methylation in the intron-2 determines the expression and estrogen regulation of cystatin A in breast cancer cells.
   Dixcy Jaba Sheeba, John Mary, Girija Sikarwar, Ajay Kumar, Anil Mukund Limaye Department of BSBE, IIT Guwahati, Guwahati, 781039, Assam, India.

2. A Variational Bayes Beta Mixture Model for feature selection in DNA Methylation studies Zhanyu ma, and Andrew E. Teschendroff KTH-Royal Institute of Technology School of Electrical Engineering SE-100 44, Stockholm, Sweden.

3. A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways Tarmo Äijö, Yun Huang, Henrik Mannerström, Lukas Chavez, Ageliki Tsagaratou, Anjana Rao and Harri Lähdesmäki

4. An evaluation of statistical methods for DNA methylation microarray data analysis
   Dongmei Li1, Zidian Xie, Marc Le Pape and Timothy Dye

5. Higher order methylation features for clustering and prediction in epigenomic studies Chantriolnt-Andreas Kapourani and Guido Sanguinetti1,2

6. The EM Algorithm: A Guided Tour
   Christophe Couvreur Service de Physique Generale, Faculte Polytechnique de Mons Rue de Houdain 9, B-7000 Mons, Belgium

7. Theory and Use of the EM Algorithm
   Maya R. Gupta and Yihua Chen

8. DNA Methylation and Cancer
   MartaKulis, ManelEsteller Cancer Epigenetics and Biology Program (PEBC), The Bellvitge Institute for Biomedical Research (IDIBELL), Hospital Duran i Reynals, Avinguda Gran Via de L'Hospitalet 199-203, E-08907 L'Hospitalet de Llobregat, Barcelona, Catalonia, Spain

9. Clustering DNA methylation expressions using nonparametric beta mixture model
   Lin Zhang, Jia Meng, Hui Liu and Yufei Huang