Identification of Cancer Drivers using Classification of large DNA Methylation Dataset

Sanchi Bansal[1], Shivam Kumar[2], Shubham Pandey[3], Stuti Saxena[4]
[1,2,3,4] Student, *Bharati Vidyapeeth (Deemed To Be) University, College of Engineering, Pune, India*
Mrs. Rohini Khalkar[5]
[5]*Asst. Prof. Bharati Vidyapeeth (Deemed To Be) University, College of Engineering, Pune, India*

***Abstract*- A well-studied genetic modification is crucial to regulate the functioning of the genome, which is done with the help of DNA Methylation. Alteration of DNA plays a vital role in tumor generation (tumorigenesis) and tumor-suppression. Therefore, studying DNA methylation data may help in identifying basic molecules or elements in body that indicates the presence of cancer. DNA methylation related data available from the public is huge – and considering the high number of methylated sites (features) present in the genome – it is crucial to have a technology for efficient processing of huge datasets. With the help of big data technologies, we propose an algorithm that can apply supervised learning in the form of classification methods to datasets with large amount of features. Through iterative deletion of selected features, extraction of equivalent classification models is possible using this algorithm. The experiments will be executed on DNA methylation datasets extracted from The Cancer Genome Atlas, where we will be focusing on three types of tumors: breast, kidney, and thyroid carcinomas. Several methylated sites and their associated genes will be extracted and classification will be performed on them with accurate performance. Thereafter, we will study the performance of our algorithm and compare it with other classifiers and with existing approaches used to analyze this data i.e, a wide-spread DNA methylation analysis method based on network analysis. Finally, we will be able to efficiently compute multiple alternative classification models and extract a set of candidate genes from DNA-methylation large datasets to be further examined to determine their role in cancer.**

Index Terms- DNA methylation; machine learning; cancer; disease diagnostic predictive models; algorithm and techniques to speed up the analysis of big medical data.; classification.

I. INTRODUCTION

Tumor, or neoplasm, is a mass of tissue originated from an abnormal and uncontrolled division of eukaryotic cells. When tumoral cells begin to destroy its surrounding tissues, the tumor is malignant and it is called cancer. According to the World Health Organization(http://www.who.int/mediacentre/factsheets/fs297/en/), nearly one six of death are caused by cancer. Since it is evident these days that cancer is one of the leading causes of increase in mortality rate, it is extremely important that research to understand its mechanisms and discover new ways using latest technologies to prevent and to treat this disease is elementary to the human race. The interaction of genetic factors with external agents, like viruses, chemicals and physical mutagens, results in a complex process of transformation of healthy cells to tumoral ones. This is the reason why the importance of DNA methylation in carcinogenesis is widely recognized. One of the most intensely studied genetic modification in mammals that involves reversible covalent alterations of DNA nucleotides, is DNA Methylation. The enzyme DNA methyltransferase acts as a catalyst in the conversion of the cytosine (typically in a CpG site) to 5-methylcytosine, by adding a methyl group (CH 3) to cytosine residues in the sequence. In normal cells, this conversion results in different interaction properties that assure the proper regulation of gene expression and of gene silencing. In the haploid human genome there are around 28 million of CpG sites in methylated or demethylated state. Hyper-methylation within the gene regions may result in inactivation of tumor- suppressor genes. A large range of cancer related genes can be silenced by DNA methylation in different types of tumors. Futhermore, Hypo- methylation, induces genomic instability, and contributes to cell transformation.

Therefore, methylation corresponds to inactivity, but inactivity of a repressive factor results in stimulation. This implies that studying DNA methylation data with the intent to identify cancer drivers is a challenging task. After the employment of Next-Generation Sequencing technologies, reduction of the cost of data generation has made available an enormous amount of raw data, which has posed a challenge for us. The availability of big datasets creates problems when analyzing with the application of classical data mining and analysis algorithms. In our work, we will focus on the adoption of big data technologies for the applying classification algorithms on large DNA methylation datasets. There can be many different definitions of big data, but in our context, "Big data refers to datasets whose size is beyond the ability of traditional database software tools to capture, store, manage and analyze". This definition focuses on the technology we adopt to manage the datasets and not their specific size. It is our aim to extract a set of genes playing a role in any specific tumor by application of supervised learning methods to DNA methylation datasets with a large number of features. We want to compute many classification models containing genes by applying optimized supervised learning algorithms, like Decision Trees and Random Forests. With the help of Apache Spark MLlib, running in standalone or cluster mode, we aim to cope with performance. The largeness of the input dataset prohibits the analysis and processing in an acceptable time with non-big data technologies. A previous classification study on DNA methylation proposed MethPed, which is a tool for the identifying pediatric tumors. The classification model were built by researchers behind MethPed from DNA methylation datasets with 450 thousand of features. At first, they applied a large number of regression algorithms and selected a subset of features with the highest predictive power; then, they adopted Random Forests to build the classification model. On the other hand, we want to apply classification algorithms to the entire dataset and obtain a large number of CpG sites and their associated genomic locations. Another study described methylKit, which is an R package for the analysis of DNA methylation data. This package adopted an unsupervised machine learning technique, by working on unlabeled data. methylKit worked in-memory and, even if it was multi-threaded, its execution was limited to a single machine. On our side, we want to perform supervised machine learning on a cluster of computational nodes, and to be able to scale with the increasing dimension of input data. The algorithm proposed in our study is inspired by CAMUR and is being applied to large input datasets. CAMUR (Classifier with alternative and multiple rule-based models) is a classification method that is responsible for iteratively computing a rule-based classification model, eliminating combinations of extracted features from the input dataset, and repeating the classification until a stopping condition is verified. A set of classification models is the result of a CAMUR computation. CAMUR has already been working on RNA sequencing cancer datasets with around 20 thousand features. In this work, we design and develop an algorithm, which is a multiple tree-based classifier, to analyze DNA methylation datasets with even larger number of features. Our goal is to extract methylated sites and their related genes.

1. Methods

In our experiments on the application of big data technologies to the classification of large DNA methylation datasets, we consider three types of cancer: the Thyroid Carcinoma (THCA), the Breast Invasive Carcinoma (BRCA) and the Kidney Renal Papillary Cell Carcinoma (KIRP). We develop an algorithm to run an iterative classification algorithm in big data, with the intent to achieve an efficient supervised learning technique, and to extract multiple classification models. Then, we will test our algorithm both in a single-machine and in a Hadoop YARN cluster.

2.1 Datasets

A project started in 2005 and was maintained by the National Cancer Institute and National Human Genome Research Institute (Weinstein et al., 2013) is called The Cancer Genome Atlas (TCGA). The TCGA is a public dataset of about 2.5 petabytes widely used in scientific research. Searching "The Cancer Genome Atlas" gives more than 2,500 articles of 5 years. The genomic characterization of over 30 types of human cancer from more than 11,000 patients are contained in The TCGA datasets. The cancer genome profiles obtained from several NGS methods applied to patient tissues, like Array-based

DNA methylation sequencing, RNA sequencing, microRNA sequencing, and many others are also included in the dataset. In our work, we will be focussing on DNA methylation data. In particular, profiles obtained using the Illumina Infinium Human DNA Methylation 450 platform (HumanMethylation450), will be considered, which are capable of providing quantitative methylation measurement at CpG site level. HumanMethylation450 allows assessing the methylation status of more than 450 thousand CpG sites, producing large datasets which have to be analyzed and interpreted. Even if HumanMethylation450 datasets can be useful for large-scale DNA methylation analysis, they raise problems of efficient data processing. Thereafter, we will explore the adoption of big data technologies and infrastructures and enable the possibility of efficient application of machine learning algorithms to such large datasets. We depend on the latest TCGA data available at The Genomic Data Commons data sharing platform (https://gdc.nci.nih.gov/).

In our experiments, we will be using the beta value as an estimate of DNA methylation level. Beta value can be defined as the ratio of the methylated allele intensity and the overall intensity (i.e. the sum of methylated and unmethylated allele intensities):

where $Meth_n$ is the n^{th} methylated allele intensity, $Unmeth_n$ is the n^{th} unmethylated allele intensity, and ε is a constant offset used to regulate the beta value where both have low intensities. It is worth observation that beta value is a continue variable in the range [0, 1], where 0 means no methylation and 1 full methylation.

Table 1. Datasets used in this study

Dataset	Number of Samples	Number of Features
BRCA	807	486,512
KIRP	521	486,512
THCA	421	486,512

We will focus on three DNA methylation datasets extracted from TCGA: BRCA, THCA, and KIRP (Table 1). For each dataset, we have to filter the input data matrix to cope with missing values and exclude control cases, which is important to reduce the classification task to binary classification, having only tumoral and normal cases. The final data matrix (Table 2) has the following structure:

- Rows are used to represent samples, i.e. the profile of a patient tissue. The first row is the header and it contains the column names.
- The first column has the ID of samples. The last column is the category, specifying whether the sample is “tumoral” or “normal”.
- All other columns will represent CpG sites, and the corresponding cells will contain the beta value for the CpG site. We use the Illumina 450k manifest for knowing where a CpG site is located and which gene will correspond to it. The manifest is available on Illumina website (https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html).
- Missing values are represented with the question mark.

Table 2. Structure of the DNA methylation data matrix extracted from TCGA

Sample Id	cg13869341	..	cg0381604	class
TCGA-A7-A0DC-11	0.971644	..	0.017485	Tumoral
TCGA-BH-A0BV-11A	0.925557	..	?	Normal
TCGA-BH-A0DZ-11A	0.907020	..	0.019204	Tumoral

2.2 Supervised Learning

In our project we are using an algorithm that is iterative, which in turn extricate a set of genes/features from the cancer dataset which is a large DNA methylation cancer datasets. The first step of our project is the application of a supervised learning method, because we are using labeled dataset for iteration and testing i.e. we already know if each tissue is of ‘normal’ or ‘tumoral’ category. Using a labeled dataset as a training set and using the supervised learning algorithm infers and builds a classification model, which is simply a algorithm that will classify a sample as per supervision. We perform tests with Random Forests and Decision Trees. Then we extricate the CpG sites from the classification model and the respective genes. The list of genes from the CpG sites that will be extricated from a

classification model is one of the output of our iterative algorithm that we are going to implement in our project.

In our project the algorithm will run many iterations, and the final result will be the sum of the results of every iteration. But we are not interested in the decision model to classify new data, but interested in extrication of a list of candidate genes that may play a vital role in cancer drivers.

Decision Trees are used for recursive binary partition of the feature. It starts from the root, which will contains the entire training dataset. In Decision Trees we split the dataset into distinct nodes, where each node will be of a certain category. The final node will the prediction label of the node reached during the process. Decision Trees are easy to understand and even allow approval of the model with statistical tests. It is very easy to create a tree that will satisfy the input data. In addition Decision Trees use a greedy algorithm therefore optimal tree is mostly expected.

Random Forests solves the problem of Decision Trees, especially when used for very large datasets. In Random Forest many Decision Trees can run parallelly and they fit very perfectly with MapReduce algorithms, therefore we can use splitting technique of MapReduce by implementing it in different machines. In Random Forest there are main two points of randomness, it reduces the possibility of overfitting and over generalization too. Each tree is created from a random selection of some data of the training set. Then, during the decision process it randomly selects M features from the global set of features. Because of these reasons we are using both Decision Trees and Random Forests, but the final implementation of our project is purely based on Random Forests method of Classification Algorithm.

2.3: A multiple tree-based classifier for big biological data:

In our project we are going to use CAMUR as a supervised method for training that will be an extricate alternative and builds an equivalent classification models from a labeled dataset that we are providing as a training dataset. CAMUR uses an iterative feature elimination technique which uses the supervised RIPPER algorithm for computing a rule-based classification model and iteratively eliminating the combinations of features that will be appearing in the model from the input dataset, and performs multiple iterations of the classification until a stop condition is verified. If a feature is eliminated from the dataset, it can be again inserted in the next iteration or will be discarded forever. CAMUR has been successfully applied to RNA-sequencing data extracted from TCGA also. Datasets used in CAMUR tasks contained at most 30 thousand of samples. Therefore we are going to implement it in our DNA methylation dataset for the classification and supervised method.

In our work we propose a project, a JAVA command-line software that will effectively manage a classification of large DNA datasets. Our project will adopt big data solutions.

- Our project will based on MLlib, the Apache Spark's scalable machine learning library. Apache Spark will allow executing the algorithm on Hadoop YARN cluster, which will allow us to run parallelize machine learning task on several machines present on several locations.
- We are going to use both Decision Trees and Random Forests, but the final implementation of our project will be based on Random Forests. Random Forests adopts parallel computation, as each node of a cluster can use different tree of the forest and send the final result to a main node.
- After each iteration, our project will permanently remove all features that will appear in the computed model from the dataset. Our approach is very similar to the CAMUR loose execution mode, as removing all extracted features makes the entire process lighter and makes implementation easier. Since there are hundreds of thousands of features that will guarantees a relevant number of alternative classification models.

Our project which deals with iterative procedure will stop when the reliability of the classification model is lesser or higher than the threshold value. Both of these conditions must be already specified by the user in command-line. We will use F-measure to evaluate the accurate result of classification models. The F-measure is defined as

F - Measure = 2PR/(P+R) (2)

When both precision and recall are high then F-measure is high .Precision and recall is called in the terms of TP(true positive-the number of samples that

are assigned to a category and that belong to that category), FP(False Positive-the number of samples not belonging to a category but assigned to that category), and false negatives FN (the number of samples belonging to a category but not assigned):

$P = TP/(TP+FP)$; $R = TP/(TP+FN)$ (3)

When the iterative algorithm will stop, the software will enlist all the features that is going to appear in overall computed classification models. Since features in our project is CpG sites that are located in different regions of our body. Therefore we are going to use a mapping file for discovering the region where CpG site is located. The software will enlist candidate genes as final output of the computation. The Extracted genes will be explored and evaluated by biologists. Obviously this project can be applied to different datasets.

3. CONCLUSION

In conclusion, our project will be easily managing the large DNA dataset and building an iterative and equivalent classification models for extracting features. The algorithm we are going to use is classification which could be improved. We will currently train the classification model on 80% of input data, using 20% of data as test data. This choice was important during the designing of our project. To avoid loss of information we are using this choice otherwise we could use 100% data for input even.In addition,our project can be applied for any other type of data, including other NGS experiments and even for the bigger datasets also. Lastly, our project can be used as a component to give sense to raw data by reducing the entropy and focusing on a smaller set of dimensions.

REFERENCES

[1] Akalin,A. et al. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology, 13:R87

[2] Antonucci,I. et al. (2017) A new case of "de novo" BRCA1 mutation in a patient with early-onset breast cancer. Clin Case Rep, 5(3), 238-240

[3] Bartlett, T.E., et al. (2014) A DNA methylation network interaction measure, and detection of network oncomarkers. PloS one 9.1, e84573.

[4] Baylin, S.B., et al. (2001) Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Human molecular genetics 10.7, 687- 692.

[5] Baylin, S.B. (2005) DNA methylation and gene silencing in cancer. Nature Reviews. Clin Oncol, 2(S1), S4.

[6] Bird, A. (2002) DNA methylation patterns and epigenetic memory. Gene dev. 16(1), 6-21.

[7] Breiman, L (2001) Random Forests. Machine Learning 45(1), 5-32

[8] Cestarelli, V. et al. (2016) CAMUR: Knowledge extraction from RNA-seq cancer data through equivalent classification rules. Bioinformatics, 32(5), 697-704

[9] Cohen, W.W. (1995) Fast effective rule induction. Proceedings of the twelfth international conference of machine learning, 115-123.

[10] Danielsson, A. et al. (2015) MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. Clinical Epigenetics,