Discovering Patterns of DNA Methylation: Rule Mining with Rough Sets and Decision Trees, and Comethylation Analysis

Niu Ben[1], Qiang Yang[1], Jinyan Li[2], Shiu Chi-keung[3], and Sankar Pal[4]

[1] Department of Computer Science and Engineering, Hong Kong University of Science & Technology, Hong Kong, China
{csniuben, qyang}@cse.ust.hk
[2] Institute for Infocomm Research, Singapore
JYLi@ntu.edu.sg
[3] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
csckshiu@comp.polyu.edu.hk
[4] Indian Statistical Institute, Kolkata, India
sankar@isical.ac.in

Abstract. DNA methylation regulates the transcription of genes without changing their coding sequences. It plays a vital role in the process of embryogenesis and tumorgenesis. To gain more insights into how such epigenetic mechanism works in the human cells, we apply the two popular data mining techniques, i.e., Rough Sets, and Decision Trees, to uncover the logical rules of DNA methylation. Our results show that the Rough Sets method can generate and utilize fewer rules to fully separate the methylation dataset, whereas Decision Trees method relies on more rules but involves fewer decision variables to do the same task. We also find that some of the gene promoters are highly comethylated, demonstrating the evidence that genes are highly interactive epigenetically in human cells.

1 Introduction

DNA methylation is the epigenetic modification of eukaryotic DNA involved in various biological activities including gene silencing, X chromosome inactivation, gene imprinting, and genome defense [1]. Gene silencing mediated by DNA methylation has tight relation to the tumorgenesis and the embryogenesis in human cells. In tumor progression, the aberrant methylation of the normally unmethylated promoter CGI has been found to be associated with the transcriptional inactivation of over half of the classic tumor suppressor genes. In stem cell development, the methylation of the promoter CGI is highly involved in the maintenance of the embryonic stem cell pluripotency, and the orderly differentiation of these cells into many other cell types. Analysis of the DNA methylation patterns thus has its clinical significance to the treatment of cancers and cell therapy.

Recently, DNA methylation has aroused considerable interests in the area of computational biology and bioinformatics. The classification, clustering, and feature

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 389 – 397, 2007.

selection methods in machine learning have been successfully applied to the DNA methylation data [2-6]. In this paper, we employ two data mining techniques, Rough Sets and Decision Trees to learn the logical rules of DNA methylation. In doing so, we are able to advance our knowledge on the epigenetic mechanism of human caner cells, embryonic stem cells, and normally differentiated cells.

2 Method

2.1 Data Set

We use the recently published data from Bibikova et al. [3] in our study. Three types of cells, namely, the human embryonic stem cell, the cancer cell and the normally differentiated cell, are collected and investigated, as shown in Table I,

Table I. Sample cells used in experiment

Cell type	Name of sample cells
ES cells	BG01, BG01V, BG02, BG03; ES02, ES03; HUES7; NTERA-2; Relicell hES1; SA01, SA02, SA02.5; TE04, TE06; WA01, WA07, WA09.
Cancer cell	A431; C33A; EC109; Fet; HCE4; HCT116, HT29; LNCaP; LS174; MCF7; MDA_MB_435, MDA_MB_468; NCI_H1299, NCI_H1395, NCI_H2126, NCI_H358, NCI_H526, NCI_H69; PC3; SW480; T47D, TE3, TT, TTn.
Normal cell	Breast; Colon; Lung; Ovary; Prostate; NA06999, NA07033, NA10923, NA10924.

Totally, there are 37, 24, and 9 sample cell lines collected for each type of the cells, and 1536 CpG sites are selected from the 5' regulatory regions of 371 genes. The selected genes are chosen on the basis of their importance to the cellular behaviors, including the tumor suppressor genes, the oncognenes, and the genes that are responsible for the cell growth, the apoptosis, the DNA damage repair and the oxidative metabolism. The profiling of DNA methylation consists of three major steps, the extraction of DNA, the Bisulfite conversion of the CpG sites, and the GoldenGate assay of the methylation levels. For a specific CpG site its methylation level $\alpha \in [0,1]$ is defined as the ratio of the intensities of the fluorescent signals from the methelated (m) and the unmethylated (u) alleles,

$$\alpha = \frac{m}{m+u}. \tag{1}$$

Finally, we have 70 entries of the different types of cells. Each of them contains 1536 attributes of the CpG sites with the numerical attribute values in [0, 1].

2.2 Rough Sets Method

Rough Sets theory was first developed by Pawlak in the early 1980s [7]. It is an effective machine learning method that can be utilized to represent and reason about the imprecise and the uncertain data. We can apply Rough Sets to extract the DNA methylation rules directly from the training data set. Let A represent the set of the CpG sites. u_i^1 and u_j^2 denotes the i-th and the j-th cell sample in class C_1 and C_2, respectively. Define $f_{ij} \subseteq A$ as the set of the CpG sites whose methylation intensities in the cell sample u_i^1 and u_j^2 are not identical, i.e.,

$$f_{ij} = \left\{ a \in A \mid a(u_i^1) \neq a\left(u_j^2\right), i = 1,..., N_1, j = 1,..., N_2 \right\}, \tag{2}$$

where $a(u)$ is the function that returns the value of the attribute 'a' of sample 'u'. N_1 and N_2 is the number of samples of the two classes, C_1 and C_1. The methylation rules can be generated by evaluating the Bool function in (3).

$$f = \wedge\left(\vee f_{ij}^k\right), \tag{3}$$

where f_{ij}^k is the $k-th$ element in f_{ij}, $k = 1,...,\left|f_{ij}\right|$. $\wedge$ and $\vee$ are the Bool operators. Given some new input the methylation rules can then be used for classification.

2.3 Decision Trees

Decision Trees method, introduced by Quinlan [8], is popular for rule induction due to its mathematical simplicity and the Bayesian optimality. Let A be the set of all the CpG sites and S is the sample set of two classes C_1 and C_1. For the CpG site, 'a', $a \in A$, let $Value(a)$ represent the set of its possible values and S_v is the subset of the samples whose attribute 'a' takes the value of v, $v \in Value(a)$. The construction of the decision tree is based on the computation of Information Gain by (4),

$$Gain(S, a) = Entropy(S) - \sum_{v \in Value(a)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{4}$$

where

$$Entropy(S) = \sum_i - p_i \log_2 p_i \tag{5}$$

and p_i, $i = 1, 2$, is the conditional probability of the class C_i in the sample set S. We choose the attribute with the maximum information gain to separate the samples. Recursively, in doing so, we can derive a decision tree where all the samples are classified in the leaf node. The association rules can thus be obtained by traversing

from the root node to the leaf node of the tree, and then truncating the intermediate nodes with the Bool 'AND' operators.

3 Results

To obtain the rules that can be easily interpreted and understood by human scientists, we first discrete the data by performing clustering with the k-Means method. Five clusters are generated for each attribute to describe the different levels of methylation intensity, i.e., {Very Low, Low, Middle, High, Very High}, or {VL, L, M, H, VH} in short. Fig. 1 illustrates the result of clustering for the gene CpG site APBA2-1274.

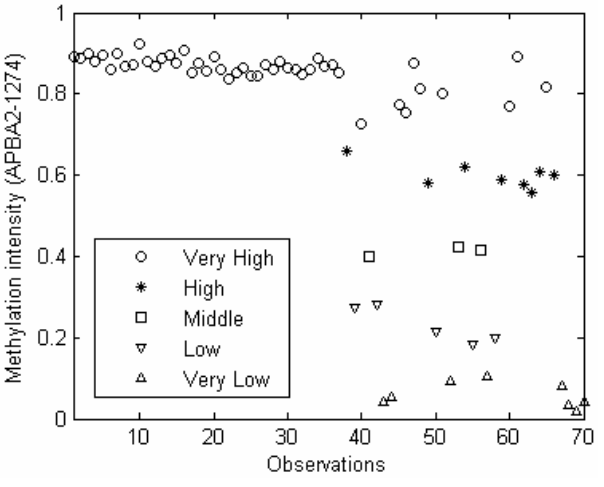

Fig. 1. Clusters generated for five language variables

3.1 Learn DNA Methylation Rules with Rough Sets

In our experiment on Rough Sets, we choose the Rosetta software [9] for implementation. We obtained three sets of methylation rules that can be utilized to

Table II. Methylation rules of embryonic stem cells

No.	Rules	Matched
1	(SMARCA3-1167=VH)&(EPO-1269=VH)&(ASC-350=VH)& (CCND2-596=VH)&(CDH3-152=VH)&(CFTR-1051=VH)	14
2	(ARHI-521=VH)&(SLC6A8-519=VH)	8
3	(ASCL2-1038=H)&(CRIP1-1227=H)	5
4	(ASCL2-1143=M)&(MOS-1474=M)&(TSC2-307=VL)	5
5	(RARRES1-893=M)&(HIC1-1081=M)&(CAPG-337=VH)	4
6	(ASCL2-1048=L)&(ASCL2-856=L)&(IRF5-1259=M)& (ASC-1350=M)	3

fully separate the different types of cells. These rules are presented in TABLE II, III and IV, respectively. They can explain the dependency relations among the genes that contribute to the differentiation of the human embryonic stem cells, the growth and migration of the cancer cells, and the normal functions of the differentiated cells.

Table III. Methylation rules of cancer cells

No.	Rules	Matched
1	(PGR-1223=VH)&(CDH3-152=VH)	10
2	(ABL1-217=VH)&(ASC-1416=VH)&(PTPRO-1357=H)	6
3	(CFTR-1097=H)&(CRIP1-1227=VH)&(CFTR-1051=VH)&(ASCL2-1038=VH)	3
4	(ASCL2-1038=M)&(HTR1B-573=M)&(ASCL2-1143=VH)&(CFTR-1097=H)	3
5	(APBA2-537=M)&(ASC-1335=M)&(ASCL2-1048=VL)&(ASCL2-856=VL)&(ATP10A-344=M)	2

Table IV. Methylation rules of normally differentiated cells

No.	Rules	Matched
1	(ASCL2-1339=VH)&(CCND2-596=VH)&(CRIP1-1227=VH)&(CYP1A1-330=VH)&(DBC1-1053=VH)&(EDNRB-1255=VH)&(EPO-1186=VH)&(EPO-1269=VH)&(GABRA5-535=VH)&(GDF10-1382=VH)&(GSTM2-1323=VH)&(HBII-52-1450=VH)&(HLA-DRA-1353=VH)	6
2	(DLC1-1012=VL)&(EPM2A-666=VL)&(F2R-473=VH)&(IL13-298=VL)	3

3.2 Induce Methylation Rules with Decision Tree

In learning the rules from Decision Trees, we use the SPASS Clementine package [10] for our implementation. The rules obtained are listed in Table V, VI, and VII, respectively.

Table V. Methylation rules of embryonic stem cells

No.	Rules	Matched
1	(PTPNS1-765=VH)&(GABRG3-1299=VH)	13
2	(PTPNS1-765=VL)&(CHGA-1371=VL)	10
3	(PTPNS1-765=M)&(HOXA11-558=M)	7
4	(PTPNS1-765=H)&(IL13-55=VH)	2
5	(PTPNS1-765=L)&(MSF-1020=M)	2
6	(PTPNS1-765=L)&(MSF-1020=VH)&(ABCB1-562=VH)	2
7	(PTPNS1-765=M)&(HOXA11-558=VH)&(PAX6-1337=VL)	1

Table VI. Methylation rules of cancer cells

No.	Rules	Matched
1	(PTPNS1-765=H)&(IL13-55=H)	7
2	(PTPNS1-765=L)&(MSF-1020=L)	6
3	(PTPNS1-765=M)&(HOXA11-558=VH)&(PAX6-1337=VH)	5
4	(PTPNS1-765=VL)&(CHGA-1371=M)	3
5	(PTPNS1-765=L)&(MSF-1020=VH)&(ABCB1-562=H)	1
6	(PTPNS1-765=M)&(HOXA11-558=VH)&(PAX6-1337=M)	1
7	(PTPNS1-765=M)&(HOXA11-558=VL)	1

Table VII. Methylation rules of normally differentiated cells

No.	Rules	Matched
1	(PTPNS1-765=VH)&(GABRG3-1299=M)	4
2	(PTPNS1-765=L)&(MSF-1020=VL)	2
3	(PTPNS1-765=VH)&(GABRG3-1299=H)	1
4	(PTPNS1-765=VL)&(CHGA-1371=L)	1
5	(PTPNS1-765=VL)&(CHGA-1371=VH)	1

In Table VIII, for comparison purpose, we show the statistics on the rules generated using the Rough Sets and the Decision Tree methods.

Table VIII. Statistics on the results of three methods

Method	Num. rules	Aver. length of rules	Num. CpGs	Num. genes
Decision Tree	19	2.1	8	8
Rough Sets	13	4.2	43	35

Decision Trees method results into 19 rules that fully summarize the methylation data, where 8 distinct CpG sites are used in building these rules. The rules have the average length of 2.1 attributes in their conditional part. The rough sets method, on the other hand, obtains 13 DNA methylation rules, six rules fewer than the decision tree method. It thus generates a more concise description of the methylation profile. But the average length of the rough sets rules is 4.2, larger than that of Decision Tree. It employs 43 distinct CpG sites of 35 genes in formulating these rules, compared with the 8 CpG sites in Decision Trees.

3.3 Co-methylation Analysis

To investigate the epigenetic interactions of genes, we compute the correlation score of the methylated CpG sites. We discover that CpG sites in the embryonic stem cells, the cancer cells and the normally differentiated cells can b highly comethylated, indicating that they are modulated by the same epigenetic process for cell function. In

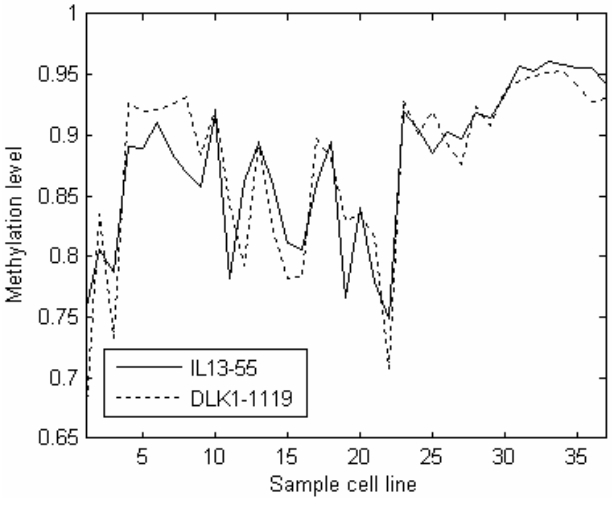

Fig. 2. Comethylation of gene IL13-55 with DLK1-1119 (embryonic stem cell)

Fig. 2, the DLK1 and the IL13 gene promoters have the comethylation score 0.88. DLK1 encodes the proteins mediating the differentiation of the B cells, while IL13 produces their growth factors. They coordinate to mobilize the immune system.

For cancer cells (see Fig. 3), the gene MYC-813 and PRKAR1A-1317 are highly comethylated with the high correlation coefficient of 0.974. The MYC gene is known as a very strong oncongene upregulated in many types of cancers. Surprisingly, the gene PRKAR1A, whose activation leads to the increased apoptosis of human B Lymphocytes, is comethylated with MYC. How this comethylation process occurs is still to be clarified.

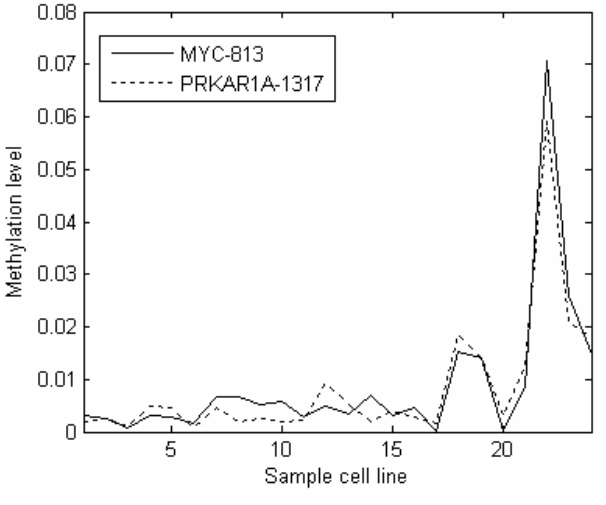

Fig. 3. Comethylation of MYC-813 with PRKAR1A-1317 (cancer cells)

Fig. 4 depicts the relation between G6PD and ELK1 in human embryonic stem cells. The comethylation score is 0.997. The G6PD produces the pentose sugars for nucleic acid synthesis, while the ELK1 regulates the nucleic acid metabolism. These two genes are involved in assembling and de-assembling the structure of the nucleic acid.

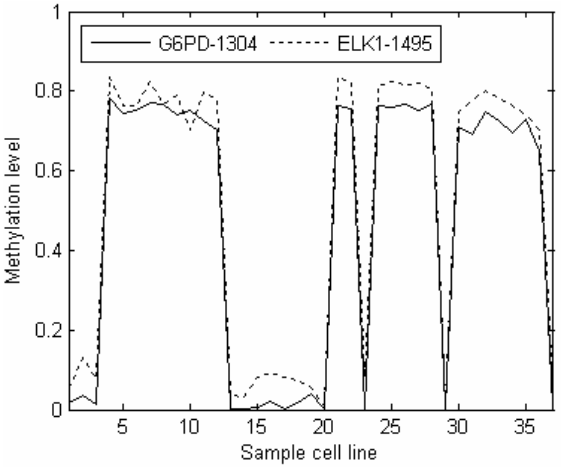

Fig. 4. Comethylation of G6PD-1304 with ELK1-1495 (embryonic stem cells)

We have performed clustering analysis on the methylation profile of all the three types of cells. The names of the genes with the comethylation coefficients higher than 0.80 are listed in the supplemental files.

4 Conclusion

We apply the two popular machine learning techniques, i.e., Rough Sets and Decision Tree, to uncover the logical rules DNA methylation in the human embryonic stem cells, cancer cells, and normally differentiated cells. Rough Sets, compared with Decision tree, generates fewer rules but involves more conditional variables to separate the three types of cells. We also demonstrate the existence of strong comethylation among the gene promoter CpG sites. Real biological experiments should be carried out in the future to identify how and why such comethylation occurs in the processes of embryogenesis, tumorgenesis, and in normal cell functions.

References

1. Jaenisch, R., Bird, A.: Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals. Nature Genetics 33 Suppl., 245–254 (2003)
2. Fabian, M., Peter, A., Alexander, O., Christian, P.: Feature Selection for DNA Methylation based Cancer Classification. Bioinformatics 17(90001), S157–S164 (2001)

3. Bibikova, M., et al.: Human Embryonic Stem Cells Have a Unique Epigenetic Signature., Genome Research, online article (August 2006)
4. Bhasin, M., Zhang, H., Reinherz, E., Reche, P.A.: Prediction of Methylated CpGs in DNA Sequences Using a Support Vector Machine. FEBS Letters 579, 4302–4308 (2005)
5. Marjoram, P., Chang, J., Laird, P.W., Siegmund, K.D.: Cluster Analysis for DNA Methylation Profiles Having a Detection Threshold. BMC Bioinformatics 7, 361 (2006)
6. Das, R., et al.: Computational Prediction of Methylation Status in Human Genomic Sequences. PNAS 103(28), 10713–10716 (2006)
7. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: Probabilistic versus Deterministic Approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
8. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1, 81–106 (1986)
9. Rosetta software, http://rosetta.lcb.uu.se/
10. SPASS Clementine software, http://www.spss.com/clementine/