Journal of
Proteomics & Bioinformatics

Zhang et al., J Proteomics Bioinform 2015, S9
DOI: 10.4172/0974-276X.S9-007

Research Article Open Access

Aberrantly Methylated CpG Island Detection in Colon Cancer

Qianwen Zhang[1], Hongwei Wu[1] and Hao Zheng[2*]

[1]Yun Nan University, Kunming, China
[2]Georgia Institute of Technology, Atlanta, GA, USA

Abstract

DNA methylation is a type of epigenetic modification which involves the addition of a methyl group to DNA via DNA methyltransferase (DNMT). In eukaryotic cells, the DNA methylation occurs at the 5' carbon position of the cytosine residue in the cytosine guanine (CpG) dinucleotide context. DNA methylation is crucial to normal organismal development, which includes cellular differentiation, genomic imprinting, X-chromosome inactivation, suppression of retrovirus transcription, etc. In addition, an accumulating volume of evidence has suggested that aberrant DNA methylation is a major contributor to the onset and progression of cancers. In this paper, we build a computational model to identify those CpG islands that are methylated in colon cancer but unmethylated in normal cells. We develop a highly accurate prediction model for those CpG islands whose methylation differentiation is related to colon cancer and evaluate these models through extensive cross-validation and generalization testing experiments.

Keywords: DNA methylation; Colon cancer; Epigenetics

Introduction

In mammals, DNA methylation is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis. Methylation of C residues spontaneously deaminates to form T residues over time; hence CpG dinucleotides steadily deaminate to TpG dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome (they occur at only 21% of the expected frequency) [1-3]. DNA methylation plays an important role in the development of cancers. A large body of evidence has demonstrated that genes with high levels of 5-methylcytosine in their promoter region are transcriptionally silent. DNA methylation can possibly affect the transcription of genes in two ways. First, the methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene, [4] and secondly, likely more important, methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs). The latter can mediates the transcriptional silencing of hypermethylated genes in cancer. There are two kinds of abnormal methylation (hypermethylation and hypomethylation) that are associated with a large number of human malignancies. Hypermethylation typically occurs at CpG islands in the promoter regions, where unmethylated CpGs are grouping in clusters. Hypermethylation is often associated with gene inactivation [5]. On the other hand, hypomethylation, in general, is linked to chromosomal instability and loss of imprinting [6].

Data Sets

Numerous databases have been constructed to archive DNA methylation profiles obtained through biochemical experiments, and to link such information with various genotypic and phenotypic information. Among these databases, Meth Cancer DB, PubMeth, and Methy Cancer contain cancer-related methylation information. Meth Cancer DB contains the data collected from over 300 resources about cancer-related aberrant CpG methylation. It focuses on the CpG islands around genes (currently covering 2,199 genes) and experimental designs such as diagnosis and prognosis [7]. PubMeth (http://mit.lifescience.ntu.edu.tw/) is based on literature search, and contains over 440 genes that are reported to be methylated in over 43 cancer types [8]. It concentrates on methylation frequency of genes in cancer samples without systematically distinguishing between cancer subphenotypes. MethyCancer integrates data from public resources (e.g., Meth DB and Human Epigenome Project (HEP)) and from data produced from China's Cancer Epigenome Project. It currently contains over 485 annotated cancer genes with methylation data from 511 cancer types [9].

In addition, mPod contains the genome-wide DNA methylation profiles in 16 normal tissues/cell types that were obtained by using the MeDIP-chip technology accompanied with bioinformatics processing. Scientists based DNA methylation profiling strategy on methylated DNA immunoprecipitation (MeDIP), a recently developed technique,which utilizes a monoclonal antibody against 5-methylcytosine to enrich for the methylated fraction of a genomic DNA sample [10,11]. Me DIP combined with microarrays is a powerful approach for DNA methylation profiling [10-13]. Recently, we have generated reference human genome-wide DNA methylation profiles for 13 normal somatic tissues, placenta, sperm, and the GM06990 immortalized cell line [14]. Several genome-wide studies show that DNA methylation profiles in mammals are tissue specific [8,15-18], and have performed the most comprehensive genome-wide study of human tissue-specific differentially methylated regions (tDMRs) and approximately 18% of the genomic regions were classified as tissue-specific differentially methylated regions (tDMRs) [16]. In promoter regions, there is a bimodal distribution of observed/expected CpG densities (CpGo/e) [17-19], whose population corresponds to CpG islands (CGIs). The recent study by Weber et al. [19] shows that a rather complex correlation between CpG-poor promoter methylation and gene expression-certain promoters with few CpGs were shown to be active and methylated, whereas other promoters of that group can be unmethylated when active. On the other hand, we observed that the populations of unmethylated non-promoter CGIs (CpGo/e

***Corresponding author:** Hao Zheng, Georgia Institute of Technology, Atlanta, GA, USA, Tel: 404-894-2901; E-mail: billzheng@sjtu.org (or) hzheng7@gatech.edu

Received January 29, 2015; **Accepted** February 17, 2015; **Published** February 23, 2015

Citation: Zhang Q, Wu H, Zheng H (2015) Aberrantly Methylated CpG Island Detection in Colon Cancer. J Proteomics Bioinform S9: 007. doi:10.4172/0974-276X. S9-007

>0.6) in the various nonpromoter categories have a strikingly similar "bell-shaped" distribution to the unmethylated CGI-promoter population. Compared with the promoter-CGIs, non-promoter CGIs are constitutively unmethylated [20].

Methods

We build a computational model to identify those CpG islands that are methylated in colon cancer but unmethylated in normal cells, which we call cancer-related differentially methylated (CRDM) CpG islands hereafter. We first combine multiple information resources to form the training data set that consists of the CRDM CpG islands and consistently unmethylated CpG islands (e.g. mpod). We then evaluate the discriminative power of various CpG island features, and obtain a lower-dimensional but more informative feature space through principal component analysis. Finally, we develop a prediction models for those CpG islands whose methylation differentiation is related to colon cancer, and evaluate these models through extensive cross-validation and generalization testing experiments.

A key step for building computational predictive models is to select features. For the prediction of DNA methylation status, our and others' previous experiences have shown that both the genetic and epigenetic information are effective. Particularly, the genetic features include: (1) general attributes (e.g., length, observed/expected CpG ratio) of the CpG island, (2) DNA composition patterns of the CpG islands, (3) distribution patterns of the functional or conserved elements within or near the CpG island, (4) structural or physicochemical properties of the CpG island, (5) functions of the genes within or near the CpG island, and (6) the extent of conservation of the CpG island among species. And, the epigenetic features mainly regard the methylation and acetylation status of the histones.

As a result, we generated 757 features using the above attribute categories. Compared to the size of our training data set (see Data Set Section), this dimension of the feature space is prohibitively high, which will potentially lead to classifier designs that are too expensive to implement or that cannotwell generalize to unseen data. Therefore, we performed a two-step feature selection procedure, where the statistical test was used to select those features that are highly correlated with the methylation (differentiation) status of CpG islands, and the principal component analysis (PCA) was used to minimize the redundancy in the features.

Two statistical tests, Chi-squared and Kolmogorov-Smirnov (KS) tests, were used to identify those features whose statistical patterns are significantly different between the positive and negative datasets. Chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is used to assess two types of comparison: tests of goodness of fit and tests of independence. The test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. The test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other. In statistics, the Kolmogorov-Smirnov test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution can be called one-sample Kolmogorov-Smirnov test, or to compare two samples named two-sample Kolmogorov-Smirnov test. Specifically, the Chi-squared tests were applied to categorical features, including the number of functional and evolutionarily conserved elements. And, the KS tests were applied to the numeric features, including CpG is land general attributes, DNA composition frequency and z-scores, average scores of functional and evolutionarily conserved elements, and the structural properties. A feature was selected if the p-value rendered by the statistical test is less than 0.05.

Although statistical tests may identify those features showing correlation with the CpG island aberrant methylation, the identified features might be inter-correlated themselves. For example, DNA sequence and structure properties are likely to be correlated, because most DNA structures are predicted based on DNA sequences. The correlation between features makes the feature space unnecessarily high-dimensional. To minimize the redundancy in the features, we performed the PCA on those methylation-related features that were selected via the above statistical tests. The PCA uses an orthogonal transformation to convert a set of values of possibly correlated dimensions into a set of values of uncorrelated dimensions called principal components. After PCA transformation, the feature components are completely decorrelated and the information contained in the original feature space before the transformation is maximally retained in the first several number of components of the new feature space. Therefore, by keeping only the first several components of the new feature space, most of the information can still be retained while the redundancy in the feature collection is greatly removed and the dimensionality of the feature space is greatly reduced. For PCA to work properly, we subtracted the mean from each of the feature dimensions.

Results and Discussions

We investigated 177 genes aberrantly methylated in colon cancer, treating the housekeeping genes as the control. Through statistical tests, we identified 75 features having different distribution between aberrantly methylated and constantly unmethylated CpG islands. These 75 features include one CpG island specific attribute, 58 DNA composition features, eight DNA structure features, six TFBS-related features, and two evolutionarily conserved element-related features. By using the first 40 principal components that retains 99.99% of the variance, our support vector machine based classifier can reach 99% specificity, 92% sensitivity, and 92% accuracy in distinguishing the aberrantly methylated in colon cancer from the constantly unmethylated CpG islands. Figure 1 shows the histogram of the predicted scores of all CpG islands for the potential of aberrant methylation in colon cancer.

Genes with top predicted scores are HOXA3, SLITRK1, FEZF2, DLX5, and FOXD2. We further did literature search to confirm the relationship of these genes with colon cancer. Extensive literature search demonstrates that some of these genes are highly associated

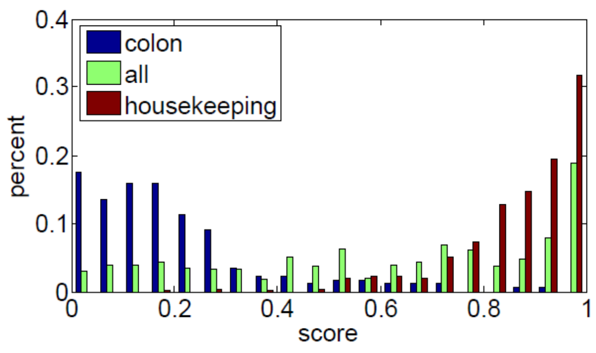

Figure 1: Histogram of the predicted scores of all CpG islands for the potential of aberrant methylation in colon cancer.

with cancer. Genomic abnormalities leading to colorectal cancer (CRC) include somatic events causing copy number aberrations (CNAs) as well as copy neutral manifestations such as loss of heterozygosity (LOH) and uniparental disomy (UPD).Combining GISTIC (Genomic Identification of Significant Targets in Cancer) ranking with functional analyses and degree of loss/gain, scientists identified eight genes in regions of significant gain, including HOXA3, as novel genes in their association with CRC [21-23]. In addition to CRC, scientists also observed that HOXA3 gene expressed in majority of breast/prostate cancer cell line cells [24] and ovarian cancer cell lines, including SK-OV3, TOV-21G, SW 626 and OV-90 [25].

Recently, several reports have indicated that single nucleotide polymorphisms (SNPs) in microRNA-target sites associate with cancer risk, treatment response and outcome [26]. miRNA target site mutations may affect function and result in cancer susceptibility [27] and have also been shown to be of potential importance for human disease as a mutation in a putative miR-189 binding-site in human SLITRK1 may be linked to Tourette's syndrome [28].

Studies also have shown that the aberrant expression of transcription factor DLX5 is also involved in some human malignancies [29] and Dlx5 can act as an oncogene in lymphomas and lung cancers [30]. DLX5 mRNA is abundantly expressed in many cancer cell lines derived from malignant tissues of breast, brain, lung, skin and ovary, but expression of DLX5 was low or undetectable in tumor cells from patients with or colorectal, prostate and kidney cancers [31]. The over expression of the DLX5 gene in mammalian cells stimulates cell proliferation [32] by regulating the expression of MYC, which can regulates transcription of numerous target genes involved in tumorigenesis [30] and the over expression can be observed in endometrial carcinoma, non-small cell lung cancer (NSCLC) and small cell lung cancer [29]. Moreover, Knockdown of DLX5 in xenografts of human ovarian cancer cells resulted in markedly diminished tumor size. In addition, DLX5 was found to cooperate with HRAS in the transformation of human ovarian surface epithelial cells. These data suggest that DLX5 plays a significant role in the pathogenesis of some ovarian cancers [31]. The knockdown of the DLX5 expression using siRNA results in the arrest of cell proliferation [32], DLX5 has a direct effect on the expression of proto-oncogene c-myc. They allow us to regard DLX5 as a promising target for which specific ligands that have the properties of oncogenesis inhibitors can be found [33]. Thus, these top predicted genes can be prioritized for molecular biologists in designing wet-lab experiments for the epigenetic association with cancer.

Conclusions

We investigated patterns indicative of methylation variation in normal tissues versus cancerous tissues. We performed the analysis of aberrant methylation in a colon cancer. We correlated various features with cancer related aberrant methylation through various statistical tests and machine learning. We also used our predictive models to all promoter CpG islands in human genome to prioritize potentially aberrantly methylated genes in cancer. Genes with top predicted scores are generated to facilitate web lab validation.

Acknowledgement

The authors would like to thank Dayu Yang for his discussions and insights.

References

1. Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321: 209-213.
2. Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP (1998) Alterations in DNA methylation: a fundamental aspect of neoplasia. Adv Cancer Res 72: 141-196.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
4. Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, et al. (2010) Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics 11: 519.
5. Zhang FF, Cardarelli R, Carroll J, Zhang S, Fulda KG, et al. (2011) Physical activity and global genomic DNA methylation in a cancer-free population. Epigenetics 6: 293-299.
6. Daura-Oller E, Cabre M, Montero MA, Paternain JL, Romeu A (2009) Specific gene hypomethylation and cancer: new insights into coding region feature trends. Bioinformation 3: 340-343.
7. Lauss M, Visne I, Weinhaeusel A, Vierlinger K, Noehammer C, et al. (2008) MethCancerDB--aberrant DNA methylation in human cancer. Br J Cancer 98: 816-817.
8. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, et al. (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol 2: e405.
9. He X, Chang S, Zhang J, Zhao Q, Xiang H, et al. (2008) MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res 36: D836-841.
10. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, et al. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853-862.
11. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, et al. (2006) Evidence for an instructive mechanism of *de novo* methylation in cancer cells. Nat Genet 38: 149-153.
12. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189-1201.
13. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39: 61-69.
14. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
15. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 38: 1378-1385.
16. Khulan B, Thompson RF, Ye K, Fazzari MJ, Suzuki M, et al. (2006) Comparative isoschizomer profiling of cytosine methylation: the HELP assay. Genome Res 16: 1046-1055.
17. Kitamura E, Igarashi J, Morohashi A, Hida N, Oinuma T, et al. (2007) Analysis of tissue-specific differentially methylated regions (TDMs) in humans. Genomics 89: 326-337.
18. Illingworth R, Kerr A, Desousa D, Jørgensen H, Ellis P, et al. (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 6: e22.
19. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99: 3740-3745.
20. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412-1417.
21. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457-466.
22. Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, et al. (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res 18: 1518-1529.
23. Eldai H, Periyasamy, Al Qarni S, Al Rodayyan M, Mustafa SM, et al. (2013) Novel Genes Associated with Colorectal Cancer Are Revealed by High Resolution Cytogenetic Analysis in a Patient Specific Manner. PLoS ONE 8: e76251.

24. Yamamoto M, Cid E, Bru S, Yamamoto F (2011) Rare and Frequent Promoter Methylation, Respectively, of TSHZ2 and 3 Genes That Are Both Downregulated in Expression in Breast and Prostate Cancers. PLoS ON3 6: e17149.

25. Hong JH, Lee JK, Park JJ, Lee NW, Lee KW, et al. (2010) Expression pattern of the class I homeobox genes in ovarian carcinoma. J Gynecol Oncol 21: 29-37.

26. Salzman DW, Weidhaas JB (2013) SNPing cancer in the bud: microRNA and microRNA-target site polymorphisms as diagnostic and prognostic biomarkers in cancer. Pharmacol Ther 137: 55-63.

27. Slaby O, Bienertova-Vasku J, Svoboda M, Vyzula R (2012) Genetic polymorphisms and microRNAs: new direction in molecular epidemiology of solid cancer. J Cell Mol Med 16: 8-21.

28. Blenkiron C, Miska EA (2007) miRNAs in cancer: approaches, aetiology, diagnostics and therapy. Hum Mol Genet 16 Spec No 1: R106-113.

29. Pedersen N, Mortensen S, Sørensen SB, Pedersen MW, Rieneck K, et al. (2003) Transcriptional gene expression profiling of small cell lung cancer cells. Cancer Res 63: 1943-1953.

30. Xu J, Testa JR (2009) DLX5 (distal-less homeobox 5) promotes tumor cell proliferation by transcriptionally regulating MYC. J Biol Chem 284: 20593-20601.

31. Tan Y, Cheung M, Pei J, Menges CW, Godwin AK, et al. (2010) Upregulation of DLX5 promotes ovarian cancer cell proliferation by enhancing IRS-2-AKT signaling. Cancer Res 70: 9197-9206.

32. Tan Y, Timakhov RA, Rao M, Altomare DA, Xu J, et al. (2008) A novel recurrent chromosomal inversion implicates the homeobox gene Dlx5 in T-cell lymphomas from Lck-Akt2 transgenic mice. Cancer Res 68: 1296-1302.

33. Xu J, Testa JR (2009) DLX5 (distal-less homeobox 5) promotes tumor cell proliferation by transcriptionally regulating MYC. J Biol Chem 284: 20593-20601.

This article was originally published in a special issue, **Computational Intelligence in Bioinformatics** handled by Editor(s). Dr. Jean-Christophe Nebel, Kingston University, London, UK