Gene Methylation Data – a New Challenge for Bioinformaticians?

G. Goebel[1], H. M. Müller[2], H. Fiegl[3], M. Widschwendter[3]
[1]Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Innsbruck, Austria
[2]Department of General and Transplant Surgery, Innsbruck Medical University, Innsbruck, Austria
[3]Department of Gynaecology and Obstetrics, Innsbruck Medical University, Innsbruck, Austria

Summary

Objectives: Changes in the status of *DNA methylation*, known as epigenetic alterations, are among the most common molecular alterations in human neoplasia. For the first time, we reported on the analysis of fecal DNA from patients with CRC to determine the feasibility, sensitivity and specificity of this approach. We want to present basic information about DNA methylation analysis in the context of bioinformatics, the study design and several statistical experiences with gene methylation data. Additionally we outline chances and new research questions in the field of DNA methylation.

Methods: We present current approaches to DNA methylation analysis based on one reference study. Its study design and the statistical analysis is reflected in the context of biomarker development. Finally we outline perspectives and research questions for statisticians and bioinformaticians.

Results: Identification of at least three genes as potential DNA methylation-based tumor marker genes *(SFRP2, SFRP5, PGR)*.

Conclusions: DNA methylation analysis is a rising topic in molecular genetics. Gene methylation will push the extension of biobanks to include new types of genetic data. Study design and statistical methods for the detection of methylation biomarkers must be improved. For the purpose of establishing methylation analysis as a new diagnostic/prognostic tool the adaptation of several approaches has become a challenging field of research activity.

Keywords

DNA methylation, biological tumor marker, genetic marker, colonic neoplasms, computational biology

Methods Inf Med 2005; 44: 516–9

Introduction

Changes in the status of *DNA methylation*, known as epigenetic alterations, are among the most common molecular alterations in human neoplasia. Several studies have reported the presence of methylated DNA in serum/plasma and other body fluids of patients with various types of malignancy and the absence of methylated DNA in normal control patients [1-3]. Therefore, epigenetic alterations represent important serologic markers for risk assessment and even for therapy monitoring during follow-up of cancer patients. DNA methylation-based tumor markers can be found in various types of circulating nucleic acids in the plasma or serum of cancer patients such as microsatellites, viral DNA, nucleosomes, mitochondrial DNA and cell-free mRNA.

It has been increasingly recognized over the past four to five years that the CpG islands of a large number of genes, which are mostly unmethylated in normal tissue, are methylated to varying degrees in all types of human cancers and that these tumor-specific alterations can be detected in tumor-derived DNA found in the serum/plasma or remote media of cancer patients [1, 4-8].

Despite the many unresolved questions, circulating DNA methylation changes represent one of the most promising tools for risk assessment in cancer patients. Nevertheless, it is not yet clear how to proceed to choose the most suitable prognostic or predictive DNA methylation markers out of a long list of candidate genes known to be hypermethylated in neoplasia.

From the viewpoint of bioinformatics much effort has been put on the development of new research strategies and methods in the field of gene expression analysis and proteomics, whereas most of the gene methylation studies are based on well-known biostatistical models.

DNA Methylation at a Glance [9]

The addition of a methyl group to the fifth position of cytosine within a CpG dinucleotide is called DNA methylation. This DNA modification is critical in normal organism development. Areas of genes that are rich in these CpG islands are shown to not usually be methylated in normal tissues but frequently become hypermethylated in cancer. It has been shown that, for example, about 70% of the CpG islands in the mammalian genome are methylated [10]. This hypermethylation is associated with gene silencing [10] and plays an important role in the inactivation of tumor suppressor genes. Methylation profiles differ between cancers arising in different organs [11, 12] and even between different cancer histologies from the same organs [13, 14]. Methylation profiling utilizes DNA for expression profiling and therefore these stable methylation signatures are detectable in samples obtained from a wide variety of body materials like stool, serum or sputum. For these reasons, DNA methylation analysis is expected to become a powerful tool for the early detection of cancer. Used as a high-throughput screening method in cancer diagnosis it offers the prospect of early detection and risk assessment for the future development of disease.

DNA Methylation in the Context of Bioinformatics

The National Centre for Biotechnology Information (NCBI) provides the biggest database network for molecular biology information. It develops and promotes standards for databases, data deposition and exchange, and biological nomenclature. Unfortunately no information about epigenetic data and experiments has been available via this institution up until now.

Methods and Approaches

For the first time we were able to report on the analysis of fecal DNA from patients with CRC to determine the feasibility, sensitivity and specificity of potential DNA methylation tumor markers [1]. The study was designed as a proof of principle study based on a small number of cases for the selection of potential genes and their validation within independent training and test sets. We used three independent sets of probands; one for the determination of potentially discriminating genes (gene evaluation set), one training model (training set) and one test set for validation.

Basics of DNA Methylation Analysis – an Example Study Design

The *gene evaluation set* was used to determine the most suitable epigenetic markers to identify patients with CRC. Either bowel lavage fluid collected during colonoscopy or mucus and bowel contents collected intraoperatively from nine CRC patients and ten control patients without CRC were used. After DNA isolation the methylation status of a total of 44 genes was analyzed. Preselection of these genes was based on previous studies [2, 15]. Using three different statistical methods (Mann-Whitney U test by using PMR (Percentage of fully Methylated Reference, see below) values, chi-squared contingency test and PAM (Prediction Analysis for Microarrays)) ten genes showed the greatest potential for identifying CRC patients.

Next, stool samples from a total of 53 endoscopically diagnosed healthy controls, 12 patients with histologically diagnosed adenoma, 11 patients undergoing control endoscopy during colorectal cancer follow-up and 31 CRC patients were collected. The patients themselves collected the stool, either the day before colonoscopy (first stool during the bowel preparation) or in the hospital the day before surgical intervention. Before starting the DNA isolation procedure we excluded all patients diagnosed with adenoma and those having had CRC to gain clearly defined groups of patients. Next, we determined two independent age-matched sets of patients (*training* and *test set*). Due to varying amounts and consistency of the stool collected by the patients and possible degradation of DNA during self-collection, all samples were checked for their DNA content. Finally we were able to isolate DNA from 26 endoscopically diagnosed healthy controls and 23 CRC patients. All people performing the isolation procedure and MethyLight analyses were blinded to the disease status of patients.

Data Analysis

DNA methylation was measured using MethyLight, a quantitative real-time PCR technique [16]. The result measurement is the percentage of methylated reference (PMR). The distribution of PMR values follows a mixture of discrete and continuous observations. PMR values generally fall between 0 and 100. Due to fluctuations in real-time PCR amplifications or incomplete methylation of the reference genes, which are used for normalization, outliers up to 500 can appear. PMR data can also be treated as dichotomous data (methylated/unmethylated) which offers the opportunity for categorical data analysis.

As in many other contexts of biomarker detection, DNA methylation measurements are used to address three different goals [17]:

- to identify genes that are differently methylated across subgroups;
- to identify profiles that predict known disease classes;
- to identify profiles that suggest novel subgroups of disease or loci that are cohesive but distinct from one another.

Features (genes) that are differently methylated across groups of patients have to be treated very carefully due to multiple testing issues and small sample sizes – as in our case. Mostly one (or a combination) of the following strategies can be applied:

- Multiple Test Correction
- splitting the data sets into training/test sets or use of cross validation
- bootstrap

For the colon data we used a combination of two univariate standard methods (Chi^2/Mann-Whitney) and PAM. PAM was originally developed by Tibshirani [18] with respect to microarray analysis and is also called "nearest shrunken centroid method". Briefly the method provides a ranked list of significant genes characterizing each diagnostic class.

Due to limited lab resources, the workload of a methylation study must be balanced between the first step (gene selection) and the following steps (statistical modeling, validation). In order to establish a (validated) multivariate class prediction model based on the selected genes, the sample size of training/test sets must be chosen according to the number of selected genes. This makes a detailed study design very difficult, as the result of the first step (the number of selected genes) can initially only be estimated.

Due to problems incurred during the process of DNA isolation from the materials of the training/test set, the sample size in this situation inhibited the establishment of a sophisticated multivariate statistical model.

Results

The methylation status of the ten genes identified in the gene evaluation set *(SFRP1, SFRP2, SFRP5, TFF1, PGR, IGFB2, CALCA, CDH13, TWIST, MYOD1)* was validated in the fecal DNA of the patients (n = 10) and controls (n = 13), repre-

senting the predetermined *training set.* We found statistically significant differences in DNA methylation at a given gene locus for *SFRP2, SFRP5, PGR, CALCA, IGFB2* ($p = 0.003$, 0.04, 0.03, 0.019 and 0.015, respectively (MW-U)) in fecal DNA of CRC patients as compared to healthy controls; 9/10 and 3/13 patients with and without CRC, respectively, had methylated *SFRP2* in their fecal DNA (sensitivity of 90% [CI 56%; 100%] and specificity of 77% [CI 46%; 95%]). These findings were validated in the fecal DNA of the test set. Using PMR values, three genes were still differently methylated between patients with and without CRC ($p = 0.017$, 0.017 and 0.047 for *SFRP2, SFRP5* and *PGR*, respectively (MW-U)); 10/13 and 3/13 patients with and without CRC, respectively, showed *SFRP2* to be methylated in their fecal DNA (sensitivity of 77% [CI 46%; 95%] and specificity of 77% [CI 46%; 95%]).

Discussion

Reflection of Study Design and Statistical Approach

This proof of principle study aimed to clarify whether it is possible to use methylation changes in fecal DNA isolated from stool samples as a screening tool for CRC. It can be categorized as the initial part of phase 1 in the development of biomarkers for tumor screening [19]. Although the results show promising discriminating genes, further evaluation of the markers and an extension of the number of subjects will be needed, as an assessment of associated factors such as stage, histology, grade etc. is also needed. Additionally no information about sex, age or smoking behaviour has been included either in this study or in many other recent methylation studies.

Data Structure and Distribution of Methylation Data

From the statistical viewpoint several issues must be considered: data structure and statistical distributions of gene methylation data constitute an upcoming topic in bioinformatics. The analysis of methylation data from other studies showed several associations of DNA methylation measurements with other parameters (e.g. age of patients, age of DNA samples, type of lab-analysis method) (publication in preparation). The lack of power using categorical or non-parametric methods can be filled with more information about the data structure of PMR values from previous studies. For this reason, a comprehensive worldwide methylation database would make an important impact.

New Statistical Approaches

Recently it was shown that using model-based approaches for feature selection and clustering of methylation data works very well with several mixture models [9]. The best result has been achieved with a Bernoulli-lognormal model which uses log-transformed positive measurements.

Computer simulations have been proposed for several issues in recent publications [9, 13, 19]. They can be used for the choice of sample sizes, to gather knowledge about the distributions and also for the selection of appropriate models for further studies. The most promising approaches for the analysis of methylation data seem to be support vector machines (SVM) [13] and adapted clustering methods [9], however they have only been performed in simulation studies based on existing data. Facing dozens of publications about the adaptation of these methods to the analysis of microarray data and proteomics, the adaptation to epigenetic data structures and issues might trigger a wave of ideas and contributions in the future.

Integration of DNA Methylation Data into Biobanking Databases

The existing worldwide methylation database (www.methdb.net) only contains a small set of data and profiles [20, 21] and is not linked to all related recent publications, in contrast to Entrez databases. A Pubmed-search at the end of October 2004 with the expression "*Methylation AND DNA*" (limited to human beings) yielded 682 publications in 2001, 847 in 2002, 1033 in 2003 and 600 until the end of October 2004. The increasing number of publications about (human) DNA methylation should motivate bioinformaticians to expand their focus in research and teaching to DNA methylation problems. The research topics will not only cover statistical questions but also problems in the fields of data modelling and storage, knowledge representation etc. Based on the very promising results of several recent studies the comprehensive development of new approaches may boost the impact of this issue within the next years.

Conclusions

DNA methylation analysis constitutes a promising, high-throughput method for rapid screening tests for diagnosis and prognosis of cancer patients. It can aid in building epigenomic cancer profiles and can be linked to epidemiological questions due to its presumed interference with drugs and/or lifestyle. Advanced methods like SVM, bootstrap etc. may help researchers in designing more sophisticated experimental designs. For this reason methylation data are a challenging new field of research for bioinformaticians regarding statistical, machine learning and epigenomic questions.

References

1. Muller HM, Oberwalder M, Fiegl H, Morandell M, Goebel G, Zitt M, et al. Methylation changes in faecal DNA: a marker for colorectal cancer screening? Lancet 2004; 363 (9417): 1283-5.
2. Muller HM, Widschwendter A, Fiegl H, Ivarsson L, Goebel G, Perkmann E et al. DNA methylation in serum of breast cancer patients: an independent prognostic marker. Cancer Res 2003; 63 (22): 7641-5.
3. Widschwendter M, Jiang G, Woods C, Muller HM, Fiegl H, Goebel G, et al. DNA hypomethylation and ovarian cancer biology. Cancer Res 2004; 64 (13): 4472-80.
4. Fiegl H, Gattringer C, Widschwendter A, Schneitter A, Ramoni A, Sarlay D, et al. Methylated DNA collected by tampons – a new tool to detect endometrial cancer. Cancer Epidemiol Biomarkers Prev 2004; 13 (5): 882-8.

5. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer 2003; 3 (4): 253-66.
6. Widschwendter A, Muller HM, Fiegl H, Ivarsson L, Wiedemair A, Muller-Holzner E, et al. DNA methylation in serum and tumors of cervical cancer patients. Clin Cancer Res 2004; 10 (2): 565-71.
7. Muller HM, Ivarsson L, Schrocksnadel H, Fiegl H, Widschwendter A, Goebel G, et al. DNA methylation changes in sera of women in early pregnancy are similar to those in advanced breast cancer patients. Clin Chem 2004; 50 (6): 1065-8.
8. Muller HM, Widschwendter A, Fiegl H, Goebel G, Wiedemair A, Muller-Holzner E, et al. A DNA methylation pattern similar to normal tissue is associated with better prognosis in human cervical cancer. Cancer Lett 2004; 209 (2): 231-6.
9. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics 2004; 20 (12): 1896-904.
10. Jones PA, Laird PW. Cancer epigenetics comes of age. Nat Genet 1999; 21 (2):163-7.
11. Costello JF, Fruhwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet 2000; 24 (2): 132-8.
12. Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. Cancer Res 2001; 61 (8): 3225-9.
13. Model F, Adorjan P, Olek A, Piepenbrock C. Feature selection for DNA methylation based cancer classification. Bioinformatics 2001; 17 (Suppl 1): S157-S164.
14. Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, et al. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomarkers Prev 2002; 11 (3): 291-7.
15. Suzuki H, Gabrielson E, Chen W, Anbazhagan R, van Engeland M, Weijenberg MP, et al. A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. Nat Genet 2002; 31 (2): 141-9.
16. Trinh BN, Long TI, Laird PW. DNA methylation analysis by MethyLight technology. Methods 2001; 25 (4): 456-62.
17. Siegmund KD, Laird PW. Analysis of complex methylation data. Methods 2002; 27 (2): 170-8.
18. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002; 99 (10): 6567-72.
19. Sullivan PM, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001; 93 (14): 1054-61.
20. Amoreira C, Hindermann W, Grunau C. An improved version of the DNA methylation database (MethDB). Nucleic Acids Res 2003; 31 (1): 75-7.
21. Grunau C, Renault E, Rosenthal A, Roizes G. MethDB – a public database for DNA methylation data. Nucleic Acids Res 2001; 29 (1): 270-4.

Correspondence to:
Georg Goebel, PhD
Department of Medical Statistics, Informatics and Health Economics
Innsbruck Medical University
6020 Innsbruck
Austria
E-Mail: georg.goebel@uibk.ac.at