Gene 651 (2018) 166–173

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Prediction of DNA methylation in the promoter of gene suppressor tumor

Imane Saif[a], Yassine Kasmi[a], Karam Allali[b], Moulay Mustapha Ennaji[a,*]

[a] Team of Virology, Oncology and Medical Biotechnologies, Laboratory of Virology, Microbiology, Quality and Biotechnologies/ETB, Faculty of Science sand Technologies–Mohammedia, Hassan II University of Casablanca, Morocco
[b] Laboratory of Mathematics & Applications, Faculty of Sciences and Technologies, Hassan II University of Casablanca, PO Box 146, Mohammedia 20650, Morocco

ARTICLE INFO

Keywords:
CpG island
Hidden Markov model
Personalized medicine
Suppressor gene tumor

ABSTRACT

The epigenetics methylation of cytosine is the most common epigenetic form in DNA sequences. It is highly concentrated in the promoter regions of the genes, leading to an inactivation of tumor suppressors regardless of their initial function. In this work, we aim to identify the highly methylated regions; the cytosine-phosphate-guanine (CpG) island located on the promoters and/or the first exon gene known for their key roles in the cell cycle, hence the need to study gene-gene interactions. The Frommer and hidden Markov model algorithms are used as computational methods to identify CpG islands with specificity and sensitivity up to 76% and 80%, respectively. The results obtained show, on the one hand, that the genes studied are suspected of developing hypermethylation in the promoter region of the gene involved in the case of a cancer. We then showed that the relative richness in CG results from a high level of methylation. On the other hand, we observe that the gene-gene interaction exhibits co-expression between the chosen genes. This let us to conclude that the hidden Markov model algorithm predicts more specific and valuable information about the hypermethylation in gene as a preventive and diagnostics tools for the personalized medicine; as that the tumor-suppresser-genes have relative co-expression and complementary relations which the hypermethylation affect in the samples studied in our work.

1. Introduction

Breast cancer is a very common cancer in the women population, but also diagnosed in men with a very low rate. More than 1,384,155 new cases in worldwide and about 459,000 deaths annually (Tao et al., 2015). Several factors can be accumulated during the development of this cancer, amongst others, genetic factors that are very well known and mainly related to mutation in functional genes like the mutations in BRCA1 gene. In addition to genetic factors, cancer is strongly influenced by epigenetic factors, which play important roles in the initiation and promotion of different cancers. Amongst the functions of epigenetic is the regulation gene expression through multiple procedures without changing the genetic information (Berger et al., 2009). One of the several epigenetic mechanisms, is the methylation of both gene promoter and first exon gene; this in order to inhibit the replication (Phillips, 2008).

The deoxyribonucleic acid (DNA) methylation is defined as an addition of methyl groups to the vertebrate genome by DNA methyltransferase (DNMTs). It is an epigenetic modification that affects DNA, which is super-imposed on genetic information (Duhamel, 2007). Methylation profiles are considerably modified by the nature of the targets of methylation (for example tissue differentiation) (Chaumette, 2016).

In humans and other mammals, methylation of DNA occurs specifically on carbon 5 (C5) cytosine located 5′ of guanines (Erdmman et al., 2015). The 5-methylcytosine residues are therefore located mainly in the Cytosine-phosphate-Guanine (CpG) dinucleotides (Kgatle, 2016). The mammalian genome contains approximately 3×10^7 methylated cytosines, representing 3 to 8% of the cytosine residues of whole genome. In a human somatic cell, 5-methylcytosines represent about 1% of the whole genome, and it affects only 70–80% of the CpG dinucleotides of the genome (Bianco-Miotto et al., 2016). Generally, it is located in the promoter region of the genes. Methyl sites or the CpG

Abbreviations: DNA, deoxyribonucleic acid; CpG, cytosine-phosphate-guanine; HMM, hidden Markov model; DNMTs, DNA méthyltransférase; CGI, CpG islands; bp, base pair; BRCA1, breast cancer 1; BRCA2, breast cancer 2; CHEK2, Checkpoint kinase 2; P53, tumor protein 53; PTEN, Phosphatase and TENsin homolog; CDH1, Cadherin-1; BARD1, BRCA1-associated RING domain protein 1; STK11, Serine/threonine kinase 11; PALB2, Partner and localizer of BRCA2; MGMT, 6-*O*-méthylguanine-ADN méthyltransférase; ATM, ataxia telangiectasia mutated; RAD51C, RAD51 Paralog C; BRIP1, BRCA1 Interacting Protein C-Terminal Helicase 1; MRE11A, Meiotic Recombination 11-Like Protein A; NBN, Nibrin; RAD50, DNA repair protein RAD50; Ratio CpGo/e, ratio cytosine-phosphate-guanine (Observe/Expected); TN, True Negative; TP, True Positive; FN, False Negative; FP, False Positive; SP, Specificity; SE, Sensitivity; ACC, Positive Predictive Value; MCC, Correlation Coefficient

* Corresponding author at: Laboratory of Virology, Microbiology, Quality and Biotechnologies/ETB, Faculty of Sciences and Techniques – Mohammedia, Hassan II University of Casablanca, PB 146, Mohammedia 20650, Morocco.
E-mail addresses: allali@fstm.ac.ma (K. Allali), my.ennaji@univh2m.ma (M.M. Ennaji).

https://doi.org/10.1016/j.gene.2018.01.082
Received 15 June 2017; Received in revised form 23 December 2017; Accepted 25 January 2018

islands (CGI) in other word methylome regions have non-uniform distribution (Li et al., 2016). CGIs are often associated with promoters of most household genes and with many tissue-specific genes (Menon, 2016). In addition, methylation of CGIs is located in the promoter and very occurring in all cancers, this hypermethylation can be considered as a molecular tumor biomarker, which requires methylation prediction, due it is directly involvement in gene expression and therefore, maybe an early cancer detection tool (Erdmann et al., 2015).

The identification of CGI regions contributes to determine new candidates region for DNA methylation based on the number of base pairs (bp) and to explain the epigenetic causes of the cancer process. Many algorithms have been developed to identify methylate CGIs from a DNA sequence. For example, those basing on Frommer algorithm that request a $CpG_{o/e}$ (observed/extended) ratio greater than 0.6 and a G + C ratio greater than 50%. These algorithms allow to realize a sort of the interaction between the results obtained in order to determine the probably methylated CGI region in the selected sample (Gardiner-Garden and Frommer, 1987; Zhao and Han, 2009).

The hypermethylation at the promoter of tumor suppressor genes contributes directly to the cell cycle imbalance. The latter can be maintained only by regulation on both sides between the genes. They interact remotely via their products in the cell, which makes the gene-gene interaction very difficult to be understandable. Tumor suppressor genes maintain genomic integrity in order to prevent uncontrolled cell functions, any inactivation of these genes is directly implicating in the cancer initiation and progression process. The cell cycle characterized by the matchmaking of several proteins and the Yin-Yang dynamics of DNA methylation as aim procedures in the regulation of genes and cell phenotype (Zhuang et al., 2017). Amongst those proteins, we cite:

BRCA1, BRCA2 and CHEK2geneshavea main role to DNA damage in a various cancers (breast, ovary, etc.). It is located on the chromosome 17, 13 and 22 (Tung et al., 2015; Bosviel et al., 2012; Wang et al., 2010).

P53 for tumor protein 53 is a gene that regulates the cell cycle, located on chromosome 17 at band 13. The gene encoding P53 protein is damaged in half cancers in humans. It contains a large number of exons. Effectively, more than 50% of human cancers are due to a P53 gene Alteration (Marcel et al., 2011).

The PTEN gene is located on chromosome 10. It is a gene which plays a key role in tumor control and participates in the regulation of the cell division cycle by preventing cells from dividing too quickly and uncontrolled (Pappas et al., 2017). CDH1, BARD1 and STK11 are located respectively in chromosome 16, 2 and 19 (Aggerholm et al., 2006; Esteller, 2000) and they have as function the regulated cell cycle.

PALB2, MGMT, ATM, RAD51C, BRIP1, MRE11A, NBN and RAD50 are located in the chromosome 16, 10, 11, 17, 17, 11, 8 and 5 respectively and they have as biofunction DNA repair (Potapova et al., 2008; Wick et al., 2016; Begam et al., 2017; Hansmann et al., 2012; Smith et al., 2010; Allinen et al., 2002; Dumitrescu, 2012; Turner et al., 2004).

In this study, we had used the computational tools, hidden Markov model (HMM) and Frommer algorithms in order to predict CpG islands in the promoter of some genes. The choice of these tumor suppressors is made upon their various bio-functions. These genes are linked together in order to maintain the stability of the cell cycle.

The objective of this article is to locate and predict the profiles of methylation. Also, the gene-gene interaction is studied in order to make a link between the inactivation of a gene by hypermethylation and repression elsewhere.

2. Materials and methods

2.1. Data set

The sequences of promoter regions and the first exon of sixteen genes are collected from the National Bank for Biotechnology Information (NCBI) Bank Gene Database, as shown in the Table 1:

Table 1
List of suppressors genes used in this study, their locus and biofunction.

Gene	Code	Biofunction	locus	Reference
BRCA1	NM_007294.3	DNA damage	17q21.31	Tung et al., 2015
BRCA2	NM_000059.3	DNA damage	13q13.1	Bosviel et al., 2012
CHEK2	NM_001005735.1	DNA damage	22q12.1	Wang et al., 2010
P53	NM_000546.5	Apoptosis	17p13.1	Marcel et al., 2011
PTEN	NM_000314.6	Regulated cell cycle	10q23.31	Pappas et al., 2017
CDH1	NM_001317184.1	Regulate cell cycle	16q22.1	Aggerholm et al., 2006
BARD1	NM_000465.3	Regulated cell cycle	2q35	Esteller, 2000
STK11	NM_000455.4	Regulate cell cycle	19p13.3	Esteller, 2000
PALB2	NM 024675.3	DNA Repair	16p12.2	Potapova et al., 2008
MGMT	NM_002412.4	Repair DNA	10q26.3	Wick et al., 2016
ATM	NM_000051.3	DNA repair	11q22.3	Begam et al., 2017
RAD51C	NM_002876.3	DNA Repair	17q22	Hansmann et al., 2012
BRIP1	NM_032043.2	DNA Repair	17q23.2	Smith et al., 2010
MRE11A	NM_001330347.1	DNA repair	11q21	Allinen et al., 2002
NBN	NM_001024688.2	DNA repair	8q21.3	Dumitrescu, 2012
RAD50	NM_005732.3	DNA repair	5q31.1	Turner et al., 2004

2.2. Prediction of CpG islands

The determination of the CpG islands consists in checking their location statistically. Mathematical modeling is often used in form of recognition; hence their applications are useful for the prediction of methylated CpG.

Frommer algorithm (Eq. (1)) is based on statistical parameters of a region of at least 200 bp, with a proportion of G + C, referred to as "G + C content", of more than 50%, and one $CpG_{o/e}$ (Observe/Expected) ratio was greater than 0.6. Sliding windows consists of calculating the average concentration of the island with a pitch of 100 bases in length (Gardiner-Garden and Frommer, 1987).

$$\boldsymbol{CpG_{o/e}} = \frac{\boldsymbol{Number\ of\ CpG}}{(\boldsymbol{Number\ of\ C}) \times (\boldsymbol{Number\ of\ G})} \times \boldsymbol{N}, \quad (1)$$

where N is the total number of the DNA sequence.

In the current study, the Frommer algorithm was calculated via the CpG island Matlab package and CpG plot program at http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html.

The CGIs were also identified according to the HMM algorithm (Barazandeh et al., 2016). The basis of this algorithm is the stochastic modeling of bases in the genome. This algorithm assumes that each gene is divided into 2 states (CGI and baseline). The choice of the state is based on the Markov chain, allowing the localization of the CGI by steps of one bp. The results of the transition and emission probabilities were tested by the Viterbi algorithm to obtain an optimal solution in the sense of the maximum likelihood for the estimation of a discrete time sequence and the number of finite states observed randomly (Eq. (2)), as well as with the Baum-Welch algorithm to mathematically maximize the probability.

$$\boldsymbol{p_l}(i,x) = \boldsymbol{e_l}(i) \times \boldsymbol{max}\ (\boldsymbol{p_k}(j,x-1), \boldsymbol{p_{kl}}), \quad (2)$$

where $e_l(i)$ is the probability to observe element i (stands for C or G) in state l; $p_k(j, x-1)$ is the probability of the most probable path ending at position $x-1$ in state k with element j; p_{kl} is the probability of the transition from state l to state k.

The distribution assumption was tested by Poisson formula as follows: