Prediction of methylated CpGs in DNA sequences using a support vector machine

Manoj Bhasin[a,b], Hong Zhang[a], Ellis L. Reinherz[a,b], Pedro A. Reche[a,b,*]

[a] *Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA*
[b] *Department of Medicine, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA*

Received 9 May 2005; revised 21 June 2005; accepted 5 July 2005

Available online 18 July 2005

Edited by Robert B. Russell

Abstract DNA methylation plays a key role in the regulation of gene expression. The most common type of DNA modification consists of the methylation of cytosine in the CpG dinucleotide. At the present time, there is no method available for the prediction of DNA methylation sites. Therefore, in this study we have developed a support vector machine (SVM)-based method for the prediction of cytosine methylation in CpG dinucleotides. Initially a SVM module was developed from human data for the prediction of human-specific methylation sites. This module achieved a MCC and AUC of 0.501 and 0.814, respectively, when evaluated using a 5-fold cross-validation. The performance of this SVM-based module was better than the classifiers built using alternative machine learning and statistical algorithms including artificial neural networks, Bayesian statistics, and decision trees. Additional SVM modules were also developed based on mammalian- and vertebrate-specific methylation patterns. The SVM module based on human methylation patterns was used for genome-wide analysis of methylation sites. This analysis demonstrated that the percentage of methylated CpGs is higher in UTRs as compared to exonic and intronic regions of human genes. This method is available on line for public use under the name of Methylator at http://bio.dfci.harvard.edu/Methylator/.

Keywords: DNA; CpG; Methylation; Support vector machine; Prediction

1. Introduction

In vertebrates, the methylation of cytosine is the most frequent endogenous modification of DNA. This modification consists of the addition of a methyl group to carbon-5 in the pyrimidine ring of cytosine, and is mediated by specific DNA-methyl transferases [1–3]. The majority of the 5′-methylcytosine is present in 5′-CpG-3′ dinucleotides [4]. Methylation in non-CpG sequences is less frequent, though it can add up to 15–20% of total 5′-methylcytosine [5]. CpGs are present at an average of one per 80-dinucleotides throughout most parts of the genome. However, there are regions within the genome where CpGs are around five times greater than average. These regions are known as CpG islands [6] and comprise 1–2% of the genome. CpG islands have a high G + C content (greater than 50%), and a size ranging from 200 bp to several thousand base pairs [6,7]. The CpG islands are present in promoter and/or exonic regions of 50–60% of mammalian genes, and are mostly non-methylated. In contrast, CpGs outside CpG island are mostly methylated [3,8].

DNA methylation has been linked to the repression of gene expression through two main mechanisms. On the one hand, methylation of CpGs may disrupt the binding of certain transcription factors [9–11], and on the other, it promotes the binding of specific 5-methylcytosine binding proteins and other structural proteins which results in packing of the DNA into a structure that is inaccessible to transcription factors. Thus, it has been suggested that patterns of methylation may compartmentalize the genome into transcriptionally active (non-methylated) and non-active regions (methylated) [1,12,13]. DNA methylation can alter the flow of the genetic information and reprogram genome function [14], and therefore it has been recognized as a major epigenetic modification. Genomic methylation patterns in non-dividing somatic differentiated cells are generally stable and heritable. However, there are instances where methylation patterns undergo significant changes that alter the phenotype. For example, genome-wide changes in methylation patterns occur during developmental embryogenesis and in stem cell differentiation [14,15]. There is also evidence that methylation patterns change in CpG islands of gene promoters during aging [16–18]. Although the aforementioned alteration in DNA methylation patterns are physiological, aberrant patterns of methylation have been also found in various diseases [19], most notoriously in cancers [20]. Thus, in cancer cells, the usual pattern of DNA methylation is reversed, with hypermethylation of CpG islands in promoter genes and overall exonic hypomethylation [21–23].

Clearly, the determination of DNA methylation is functionally relevant. A high throughput method for experimental determination of overall DNA methylation has been successfully accomplished using gene expression data [24]. However, identification of the specific methylated sites in the DNA can only been accomplished through laborious and time consuming experiments. Therefore, computational identification of DNA methylation sites may provide a good alternative. Unfortunately, DNA methylation sites are poorly conserved,

[*]Corresponding author. Fax: +1 617 632 3351.
E-mail address: reche@research.dfci.harvard.edu (P.A. Reche).

Abbreviations: ACC, accuracy; AUC, area under curve; MCC, Matthew's correlation coefficient; SE, sensitivity; SP, specificity; RBF, radial basis function; SVM, support vector machine

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.07.002

until now defying computational identification. In this study, we have developed a method for the prediction of DNA methylation in CpG dinucleotides using support vector machine (SVM). The method can predict the human-specific DNA methylation sites with good accuracy outperforming alternative prediction algorithms including artificial neural networks (ANN). The method has been implemented online as "Methylator" at the site http://bio.dfci.harvard.edu/Methylator hosted by the Dana-Farber Cancer Institute. The development of this SVM-based method, the evaluation of its predictive performance, and its application to the prediction of methylation sites in the human genome will be discussed in the subsequent sections of the manuscript.

2. Material and methods

2.1. Datasets

Methylated as well as non-methylated CpG dinucleotide sequences were obtained from the MethDB database [25]. MethDB is a curated database of experimentally determined methylated DNA fragments (patterns). The database contains a total of 4996 methylation patterns form various sources ranging from plants to humans. For training of SVM the datasets were obtained by fragmenting the MethDB sequence patterns into overlapping fragments of fixed length (window size). Fragments with a methylated cytosine in the center were considered as positives, whereas fragments with non-methylated cytosine in the center were considered as negatives. Only unique fragments (non-redundant) were considered to avoid overtraining and biased predictions. Various datasets were generated by varying the window size from 9 to 89 nucleotides. Each nucleotide was represented using conventional binary sparse encoding. For example, adenine is represented by 10000, cytosine by 01000 and so on. The last bit is used to handle the incomplete windows in the initial and terminal part of the sequences.

2.2. Support vector machine (SVM)

SVM is a technique based on statistical learning theory quite popular in pattern recognition and regression problems. Here, SVM implementation was achieved using the freely downloadable package SVM_light (http://www.joachims.org) for non-commercial or academic use [26]. This package has options to select or define a kernel as well as varying the kernel parameters. Training sets consisted of N fragments or input vectors $\{x_1, x_2, x_3, \ldots, x_i, \ldots, x_N\}$ and known labels for each fragment $\{y_1, y_2, y_3, \ldots, y_i, \ldots, y_N\}$, indicating whether the fragment is methylated or not ($y_i\{+1, -1\}$). The x_i corresponds to the nucleotide sequence represented by a 5 binary sparse code (defined in Section 2.1). During training, a new value x is assigned to each fragment by the SVM according to Eq. (1)

$$\int (x) = \mathrm{sign}\left(\sum_{i=1}^{N} Y_i \alpha_i, k(x_i, x) + b\right), \tag{1}$$

where k is the kernel function that defines the feature space; b is the bias value, i is the number obtained by solving the quadratic programming problem that gives the maximum margin hyper plane. The trained SVM will give a score for each configuration varying between -1.5 and 1.5. The cytosine residue is assigned to be methylated if the predicted score is larger than the threshold.

2.3. Alternative machine learning and statistical algorithms

In this study, we have also approached the prediction of methylation sites using a set of statistical and machine learning techniques as alternatives to SVM. These techniques included ANN, Bayesian statistics, logistic regression, decision trees and the K-nearest neighbors based algorithms.

ANN. The ANN implementation was achieved using the freely downloadable package SNNS version 4.2 from Stuttgart University [27]. In the present study, a feed-forward back-propagation network with a single hidden layer was used. The network had an input window of 39 residues and 13 units in a single hidden layer. The data was provided to the network in sparse binary format (as described elsewhere). At the start of each simulation, the weights were initialized with random values. The training of the network was carried out using error back-propagation with a sum of square error function [28]. The magnitude of the error sum in the test and training set was monitored in each cycle of the training. The ultimate number of cycles (700 in this case) was determined when the network converges. During the testing of the network, the output of the network was compared with an arbitrarily defined cutoff value. If the output was greater than the cutoff, then that fragment was considered to be methylated, whereas if it was lower, it was considered as a non-methylated. The cutoff was set to the value where sensitivity and specificity are approximately equal (defined below).

Waikato environment for knowledge analysis (Weka). Weka is a java package providing an environment for implementation of a large number of machine learning and statistical algorithms [29]. In the present study, we have implemented four classifiers based on the following algorithms: (i) naïve bayes, (ii) logistic regression, (iii) J48 and (iv) lazy IBk. The naïve bayes is based on the Bayesian theorem which is particularly useful when the dimensionality of the input is high [30]. Logistic regression is a variation of the ordinary regression frequently used when the observed outcome is restricted to two values [31]. J48 is a classification algorithm that generates a decision tree by recursive partitioning of the data [32]. IBk is an algorithm based on K-nearest-neighbors that employs distance matrices to classify the data [33].

The data for all these classifiers was represented in attribute relation function format, consisting of the list of all instances with the attribute value for the instances (yes for methylated fragments and no for non-methylated fragments) separated by commas. Provided with a training and testing set, Weka generates a confusion matrix summarizing the classification results.

2.4. Evaluation of the predictive performance of the methylation classifiers

The performance of all classifiers was evaluated using a standard 5-fold cross-validation. In the 5-fold cross-validation, the dataset was randomly partitioned into 5 subsets. Each subset had an equal ratio of methylated and non-methylated fragments. Each classifier was trained 5 times, each time using 4 subsets for training and remaining the 5th subset for testing. In this way, 5 models were generated during cross-validation. The final prediction performance was obtained by averaging the results obtained from each model. Prediction performance was determined by measuring the threshold-dependent parameters sensitivity (SE), specificity (SP), accuracy (ACC) and Matthew's Correlation coefficient (MCC). The SE, SP, ACC and MCC parameters were calculated using Eqs. (2)–(5), respectively,

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}), \tag{2}$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}), \tag{3}$$

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TP} + \mathrm{FP} + \mathrm{FN}), \tag{4}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} * \mathrm{TN}) - (\mathrm{FN} * \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}}, \tag{5}$$

where TP are true positives (methylated sites predicted as methylated); FN are false negatives (methylated sites predicted as non-methylated); TN are true negatives (non-methylated sites predicted as non-methylated) and FP are false positives (non-methylated sites predicted as methylated).

Performance of the classifiers was also evaluated in a threshold independent manner by carrying out a relative operating characteristic (ROC) analysis. The ROC curves were generated by plotting the function SE versus 1-SP for various prediction thresholds [34]. The area under the ROC curve (AUC) provides a single measure of overall prediction accuracy. Values of AUC between 0.7 and 0.9 indicate good prediction accuracy, and above 0.9 indicate excellent prediction accuracy. Values of AUC between 0.5 and 0.7 indicate poor accuracy [34].

For the SVM-based modules, in addition to the 5-fold cross-validation, a leave one out cross-validation (LOOCV) was also used to evaluate the performance. In LOOCV, the classifier is successively generated on $n - 1$ samples and tested on the remaining one. This is repeated n times so that every sample is left out once. The performance of the modules is judged by computing the AUC value as indicated above.

3. Results and discussion

Most common DNA methylation occurs in cytosine of 5′-CpG-3′ dinucleotides. CpGs are present through the genome but their abundance is greater in regions known as CpG islands, which are usually located in the promoters and exons of most genes. Despite the key role DNA methylation plays in regulating gene expression and in reprogramming genome function, until now there is no method that can predict DNA methylation sites. This is due to fact that methylation sites patterns are very diffuse and cannot be identified by traditional pattern-based search algorithms. Therefore, in order to identify the complex signature for methylation of cytosine in CpGs, we have used SVM [35]. SVM is a successful learning technique that can outperform other machine learning techniques like ANN and *K*-nearest neighbors methods. Among the many attractive features of SVM algorithm are the absence of local minima, its speed and scalability, and its ability to condense information contained in the training set. It is known that methylation patterns vary from species to species [36]. Not surprisingly, methylated bacterial CpG islands can be recognized by the human immune system as foreign entities, thus triggering a powerful innate immune response [37]. Under these considerations, SVM-based methylation modules were implemented with consideration given to the taxonomic kinship of the sources of the methylation data. Thus, we have derived independent SVM modules from methylated patterns derived from human, mammalian and vertebrates (marked by a solid dot in Fig. 1). The mammalian data included methylation patterns from human, mouse and rat whereas vertebrates also included methylation data from other organisms like domestic chickens, etc. The datasets for training classifiers were prepared as indicated in Section 2 of the manuscript. We failed to develop classifiers from mouse or rat methylation data due to the low number of representative methylation patterns (<200).

3.1. *SVM module for the prediction of methylation sites in human*

For the prediction of human methylation sites, a SVM module was developed based on 2839 methylation patterns corresponding to DNA fragments from different tissues like liver, blood, kidney, epidermis, etc. SVM was trained with patterns of fixed length with methylated and non-methylated CpG dinucleotides in the center derived from human methylation data. In order to investigate the effect of flanking nucleotides in specifying the methylation, we trained SVM with sequence fragments (window size) varying from 9 to 89 nucleotides (see Section 2 for details). With a window size of 9 nucleotides (4 nucleotides at each side of the central cytosine) the SVM module performed very poorly. Thus, the determined AUC value was 0.56, indicative of a nearly random prediction. Increasing the window size from 9 to 19 dramatically improved the predictive performance of the module (AUC ∼0.73). Also, increasing the window size from 19 to 39 led to a significant improvement of the performance of the module, as judged by the augment of the AUC value from 0.70 to 0.82 (Fig. 2). However, further increases of the window size did not lead to any significant improvement in the ACC of the predictions. Consequently, a window size of 39 was selected as the optimal (set as default) for the prediction of cytosine methylation in CpG dinucleotides. At the default threshold (−0.4), where the SE and SP of the predictions are nearly equal, the module is able to achieve an ACC and MCC of 75% and 0.504, respectively. The module achieved an AUC value of 0.82, indicating that it is able to accurately model the methylation patterns. In this case, best performance was obtained using the polynomial kernel of sixth degree, clearly depicting the complexity of the methylation patterns that could not be captured by linear classifiers. The radial basis function (RBF) kernel with $g = 0.1$ also performed similar to the polynomial kernel. The detailed threshold dependent performance of the polynomial and RBF kernels is shown in Table 1. A complete summary of AUC value achieved with different window sizes has been shown in Fig. 2.

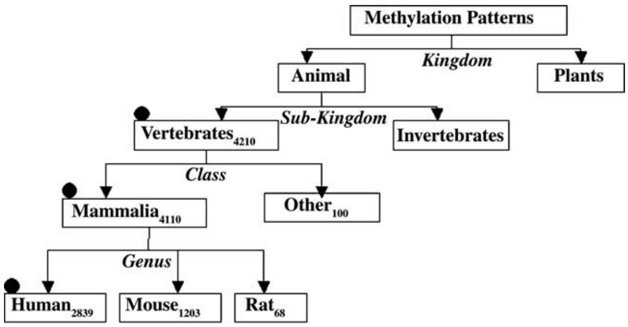


Fig. 1. Data structure for training SVM modules. Figure shows the taxonomic levels for which methylation patterns were obtained. Number indicate the methylation patterns available at the particular taxonomic levels. The SVM modules were developed for taxonomic groups marked with a solid dot.

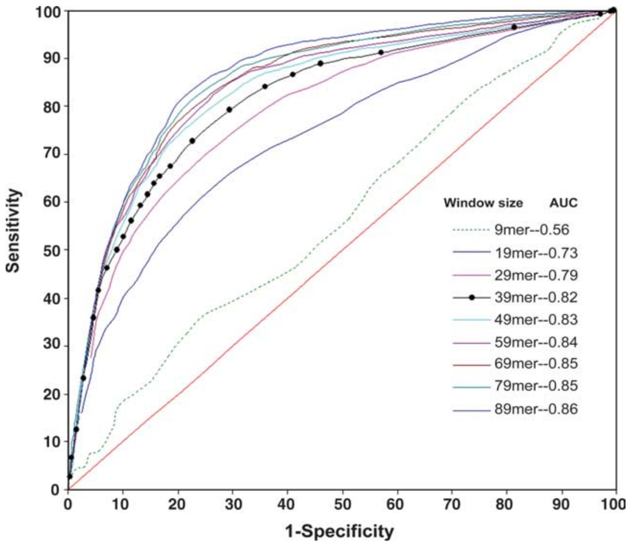


Fig. 2. Performance of SVM-modules for the prediction of CpG methylation in humans trained on different window sizes. Prediction performance of the SVM modules trained with fragments (window size) varying from 9 to 89 in 10 nucleotides increases was analyzed through a ROC analysis, and the relevant ROC curves appear plotted on the figure. The *AUC* achieved by the SVM-based models at the different window sizes is also shown in the figure. ROC curves were generated by plotting the average values of SE and 1-SP obtained through 5-fold cross-validation (see Section 2.4 for details).

3.2. SVM module for the prediction of methylation sites in mammalian data

Another SVM module was developed based on the mammalian methylation patterns that include human, mouse and rat methylation patterns. It was trained using patterns of 39-window size as it was proven previously that SVM modules based on a 39-window size performed better than other window sizes (Fig. 2). The performance of this module was evaluated using 5-fold cross-validation through threshold dependent and independent measures. This module achieved an AUC of 0.80 on the ROC analysis. The detailed performance of module in term of SE, SP, ACC and MCC is shown in Table 2.

3.3. SVM module for the prediction of methylation sites in vertebrates

Finally, another SVM module was developed from methylation patterns from vertebrates. The training and testing sets were obtained from 4210 methylation patterns corresponding to human, mouse, rat, chicken, etc. The detailed threshold dependent performance of this module is shown in Table 3. The performance of this module (AUC = 0.80) was similar to that of the SVM module based on mammalian data but lower than that of SVM modules based on human data (Fig. 3). The lower performance of the mammalian- and vertebrate-specific SVM modules as compared with the human-specific SVM module may result from the fact that methylation patterns are specific for each species. On the other hand, the small differences between the performance of the mammalian- or vertebrate-specific SVM modules and the human SVM modules are due to the fact that most of the methylation patterns in the training set corresponded to human (>60%).

3.4. Performance of the different SVM modules in LOOCV

The performance of the SVM modules (trained with fragments size of 39 residues) derived from human, mammalian and vertebrate methylation patterns was also evaluated using leave one out cross-validation (Figure S1). The performance of all modules using LOOCV technique was better than in the 5-fold cross-validation. The performance of SVM modules based on human data was 0.84, that is, 0.02 better than the performance obtained during 5-fold cross-validation. The better performance of modules is explained by the larger size of the training dataset during LOOCV as compared to 5-fold cross-validation.

3.5. Comparison of SVM-based method with other machine learning and statistical algorithms based classifiers

Recently, a few studies have shown that SVM yields better results in classifying biological data than alternative machine learning techniques [38–40]. In this study, we have analyzed the prediction of the methylation sites in human data using SVM along with other classifiers developed ex-professo and based on ANN, Bayesian statistics, logistic regression, decision trees and *K*-nearest neighbors algorithms. All these classifiers were derived from the same dataset consisting of human DNA fragments of 39 nucleotides (see Section 2). The average identity between the fragments in the datasets consisting of non-methylated fragments was 34.1%. Likewise, the identity between the methylated fragments was around 34.5%. The average identity between the methylated and non-methylated fragments was 34.2%. The performance of the classifiers based on alternative machine learning techniques is shown in Table 4. This performance was obtained using 5-fold cross-validation at a default threshold where the sensitivity and specificity are nearly equal. The accuracy of the SVM-based module was ~5% and ~7% higher than *K*-nearest neighbors and decision trees, respectively. The difference in accuracy between the SVM-based classifier and those based on ANN, logistic regression and naïve bayes classifiers was even larger (Table 4). The SVM-based module is able to recognize ~73% of the methylation sites (SE), i.e., nearly 5% higher than any of the classifier

Table 2
Performance of the mammalian SVM module

Threshold	SE	SP	ACC	MCC
−1.100	99.6	2.42	51.02	0.088
−0.900	93.6	34.36	63.98	0.346
−0.700	88.16	53.74	70.92	0.446
−0.500	83.42	62.84	73.16	0.476
−0.300	*69.42*	*75.42*	*72.42*	*0.448*
−0.100	57.04	84	70.54	0.424
0.100	52.64	87.2	69.94	0.424
0.300	46.78	89.92	68.34	0.408
0.500	40.7	92.6	66.64	0.388
0.700	30.42	95.8	63.1	0.346
0.900	10.38	98.54	54.46	0.188
1.100	1.94	99.84	50.86	0.082
1.300	0.08	20	10.04	0.01

Italicised values show the performance at default threshold.

Table 1
Performance of SVM modules trained on DNA methylation data from human

Polynomial kernel (*D* = 6)				Threshold	RBF Kernel (*g* = 0.1)			
SE	SP	ACC	MCC		SE	SP	ACC	MCC
99.88	1.02	50.42	0.062	−1.200	100	0.26	50.14	0.04
96.34	18.72	57.54	0.24	−1.000	97.38	14.32	55.86	0.21
88.9	53.98	71.42	0.458	−0.800	89.28	49.94	69.6	0.426
83.94	64.12	74.06	0.49	−0.600	85	60.32	72.66	0.468
72.74	*77.34*	*75.06*	*0.504*	−0.400	*74.46*	*73.08*	*73.76*	*0.476*
65.26	83.18	74.22	0.494	−0.200	63.72	81.38	72.56	0.458
61.52	85.56	73.54	0.486	0.000	59.26	85.06	72.18	0.46
56.08	88.56	72.3	0.472	0.200	53.86	88.24	71.04	0.45
50.02	91.1	70.54	0.452	0.400	47.22	90.66	68.92	0.42
41.58	94.38	67.98	0.422	0.600	38.28	94.16	66.24	0.392
23.24	97.16	60.2	0.304	0.800	19.74	97.68	58.72	0.278
6.64	99.32	52.98	0.158	1.000	6.72	99.24	53	0.154

Italicised values show the performance at default threshold.

Table 3
Performance of the vertebrate SVM module

Threshold	SE	SP	ACC	MCC
−1.200	99.84	0.92	50.38	0.054
−1.000	96.06	18.38	57.24	0.228
−0.800	89	53.04	71.02	0.45
−0.600	84.94	62.52	73.74	0.488
−0.400	*73.46*	*74.46*	*73.94*	*0.478*
−0.200	60.52	83.72	72.14	0.454
0.000	56.08	86.08	71.1	0.44
0.200	51.38	88.86	70.12	0.434
0.400	46.16	91.36	68.78	0.42
0.600	38.12	94.42	66.3	0.394
0.800	22.92	97.06	60	0.296
1.000	5.76	99.38	52.58	0.146
1.200	0.82	99.86	50.34	0.05

Italicised values show the performance at default threshold.

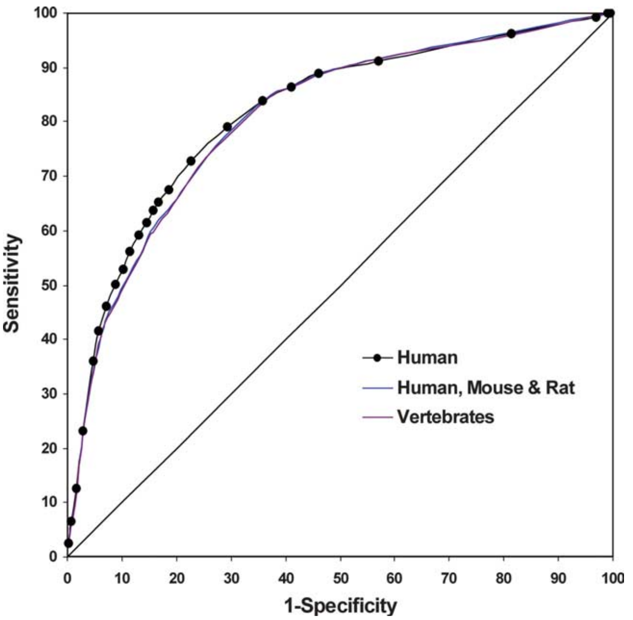


Fig. 3. Comparison of the performance of SVM modules generated from methylation patterns from different sources. Figure shows the ROC curves depicting the performance of different SVM modules based on (i) human, (ii) mammalian and (iii) vertebrate data sets.

based on alternative techniques used in this study. The decision trees and *K*-nearest neighbors classifiers were able to recognize ∼81% non-methylated sites (better than SVM) but failed to recognize the methylation sites (<59%). Thus, in our hands SVM clearly outperformed other machine learning and statistical techniques for the prediction of methylation sites in human.

3.6. Genome-wide analysis of methylation sites in human

The SVM module trained on human methylation data was applied on a dataset of 16583 known human genes obtained from NCBI. In order to examine the distribution of methylation sites in different gene structures, the module was separately applied to the UTR, exonic and intronic regions of the genes (Table 5). In brief, the results indicated that the content of the methylated CpG per 100 nucleotides is higher in UTRs (2.6), followed by exons (0.56) and introns (0.378). On the other hand, the number of CpG nucleotides per 100 nucleotides was also higher in UTRs as compared with intron and exons. Based on this study, we have developed a web server under the name of Methylator for the prediction of DNA methylation at CpG dinucleotides in humans (http://bio.dfci.harvard.edu/Methylator/). The site is meant to reduce the load on experimental biologist locating the DNA methylation sites from the genomic data, and there are plans to release a standalone version of the method in the near future. A snapshot of the home page of the server and a representative result are shown in Fig. 4.

3.7. Conclusion and limitations

In summary, we have developed an accurate method for the prediction of regular CpG methylation patterns in humans, which ought to assist biologists, reducing the load of cumbersome experiments. Furthermore, using this method, we have been able to carry out a genome wide prediction of methylation sites.

It has been noted that abnormal methylation patterns are associated with various diseases, most importantly cancer [20]. Specifically, in many tumors it has been observed that there is overmethylation – and consequent silencing – of tumor repressor genes [41]. Thus, it would be interesting to predict such abnormal patterns of methylation, as one may anticipate their potential clinical implications. That certain CpG islands associated with specific genes are more prone to overmethylation than others has already been proven from the analysis of the methylation affecting hundred of genes following the overexpression of DNMT (a specific DNA methylase) [42].

Table 5
Prediction of methylation sites in annotated human genes using the optimal SVM-based module

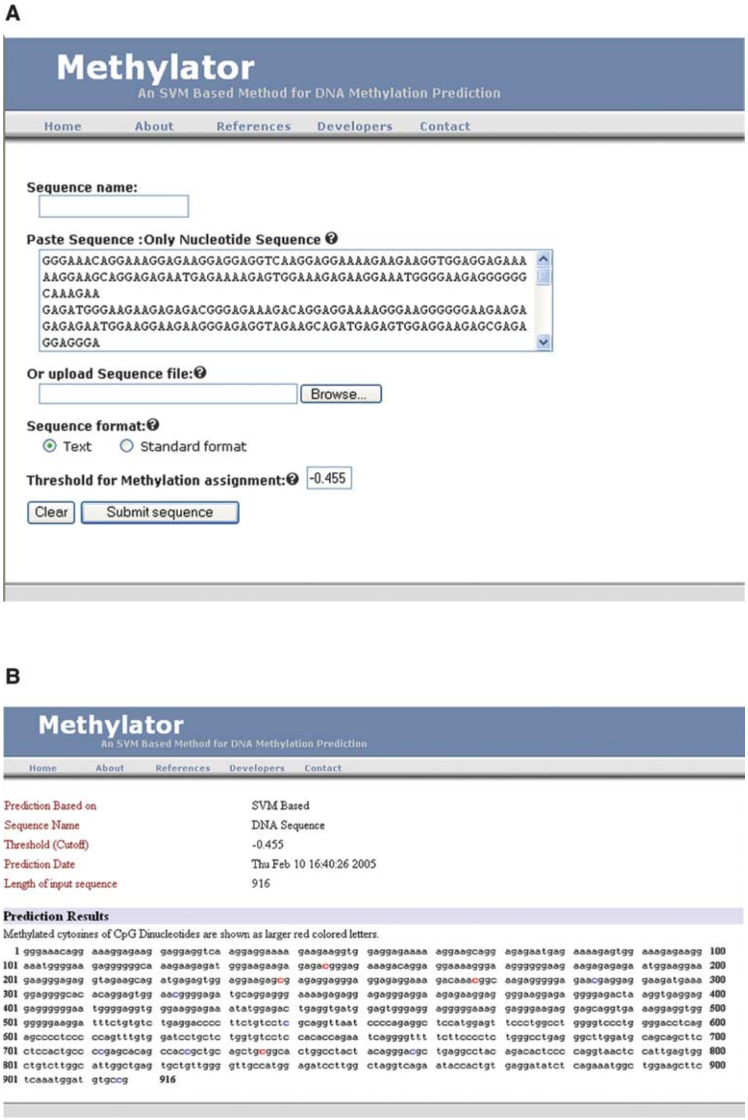
Fields	Exons	Introns	UTRs
Number	140335	106727	13148
MCpGs/100	0.560	0.378	2.6
CpGs/100	1.59	1.017	7.43

The number field indicates the total number of exons, introns and UTRs. The MCpGs/100 and CpGs/100 are methylated and total number of CpGs, respectively, per 100 base pairs.

Table 4
Performance of classifiers predicting human methylation sites

Classifier	Threshold	SE	SP	ACC	MCC
SVM	−0.400	72.74	77.34	75.06	0.504
ANN	0.30	68.02	67.74	67.88	0.3582
Naïve bayes (Bayesian statistics)	0.0	55.76	65.44	60.64	0.2132
Logistic regression	0.0	60.8	62.86	61.82	0.2368
IBk (*K*-nearest neighbors)	0.0	59.02	81.14	70.08	0.4118
J48 (decision trees)	0.0	55.58	80.96	68.3	0.378

The performance was obtained on 5-fold cross validation.



Sequence name:

Paste Sequence :Only Nucleotide Sequence

```
GGGAAACAGGAAAGGAGAAGGAGGAGGTCAAGGAGGAAAAGAAGAAGGTGGAGGAGAAA
AAGGAAGCAGGAGAGAATGAGAAAAGAGTGGAAAGAGAAGGAAATGGGGAAGAGGGGGG
CAAAGAA
GAGATGGGAAGAAGAGAGACGGGAGAAAGACAGGAGGAAAAGGGAAGGGGGGAAGAAGA
GAGAGAATGGAAGGAAGAAGGGAGAGGTAGAAGCAGATGAGAGTGGAGGAAGAGCGAGA
GGAGGGA
```

Or upload Sequence file:

Sequence format:
● Text ○ Standard format

Threshold for Methylation assignment: -0.455

Prediction Based on	SVM Based
Sequence Name	DNA Sequence
Threshold (Cutoff)	-0.455
Prediction Date	Thu Feb 10 16:40:26 2005
Length of input sequence	916

Prediction Results

Methylated cytosines of CpG Dinucleotides are shown as larger red colored letters.

```
  1 gggaaacagg aaaggagaag gaggaggtca aggaggaaaa gaagaaggtg gaggagaaaa aggaagcagg agagaatgag aaaagagtgg aaagagaagg 100
101 aaatggggaa gaggggggca aagaagagat gggaagaaga gagacgggag aaagacagga ggaaaaggga agggggaag aagagagaga atggaaggaa 200
201 gaagggagag gtagaagcag atgagagtgg aggaagagcg agaggaggga ggagaggaaa gacaaacggc aagaggggga gaacgaggag gaagatgaaa 300
301 ggaggggcac acaggagtgg aacggggaga tgcaggaggg aaaagagagg agagggagga agagaaggag gggaaggaga ggggagacta aggtgaggag 400
401 gagggggaa tggggaggtg ggaaggagaa atatggagac tgaggtgatg gagtgggagg agggggaaag gagggaagag gagcaggtga aaggaggtgg 500
501 gggggaagga tttctgtgtc tgaggacccc ttctgtcctc gcaggttaat ccccagaggc tccatggagt tccctggcct ggggtccctg gggacctcag 600
601 agcccctccc ccagtttgtg gatcctgctc tggtgtcctc cacaccagaa tcaggggttt tcttcccctc tgggcctgag ggcttggatg cagcagcttc 700
701 ctccactgcc ccgagcacag ccaccgctgc agctgcggca ctggcctact acagggacgc tgaggcctac agacactccc caggtaactc cattgagtgg 800
801 ctgtcttggc attggctgag tgctgttggg gttgccatgg agatccttgg ctaggtcaga ataccactgt gaggatatct cagaaatggc tggaagcttc 900
901 tcaaatggat gtgccg     916
```

Fig. 4. Methylator webserver. (A) Input web page. Users can enter nucleotide sequence in one of the standard formats such as GenBank, EMBL, GCG, or plain format. The method provides the option of pasting the sequence in the text area or uploading the direct sequence file. All non-standard characters except the four nucleotides bases adenine, guanine, cytosine and thymine will be ignored from the sequence. (B) Prediction results of Methylator. The methylated cytosines are highlighted in red color and bold letters. Currently the server allows only for the prediction of the DNA methylation from a single sequence, and there is a limit size of 50000 nucleotides per query. Only the best SVM module – trained on the basis of human data – has been made available in this server.

Unfortunately, these studies did not yield the location of the specific methylation sites. The availability of such abnormal patterns of methylation will be a great interest, as it will empower our method to predict gene specific overmethylation and anticipate their implications.

Acknowledgments: This manuscript was supported by NIH Grant AI57330 and the Molecular Immunology Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version at doi:10.1016/j.febslet.2005.07.002.

References

[1] Doerfler, W. (1983) DNA methylation and gene activity. Annu. Rev. Biochem. 52, 93–124.

[2] Hermann, A., Gowher, H. and Jeltsch, A. (2004) Biochemistry and biology of mammalian DNA methyltransferases. Cell. Mol. Life Sci. 61, 2571–2587.

[3] Bird, A. and Boyes, J. (1992) The essentials of DNA methylation. Cell 70, 5–8.

[4] Riggs, A.D. and Jones, P.A. (1983) 5-Methylcytosine, gene regulation, and cancer. Adv. Cancer Res. 40, 1–30.

[5] Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P. and Jaenisch, R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. USA 97, 5237–5242.

[6] Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. Nature 321, 209–213.

[7] Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA 99, 3740–3745.

[8] Bird, A., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40, 91–99.

[9] Iguchi-Ariga, S.M. and Schaffner, W. (1989) CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. 3, 612–619.

[10] Iannello, R.C., Gould, J.A., Young, J.C., Giudice, A., Medcalf, R. and Kola, I. (2000) Methylation-dependent silencing of the testis-specific Pdha-2 basal promoter occurs through selective targeting of an activating transcription factor/cAMP-responsive element-binding site. J. Biol. Chem. 275, 19603–19608.

[11] Inamdar, N.M., Ehrlich, K.C., Ehrlich, M., Iannello, R.C., Gould, J.A., Young, J.C., Giudice, A., Medcalf, R. and Kola, I. (1991) CpG methylation inhibits binding of several sequence-specific DNA-binding proteins from pea, wheat, soybean and cauliflower. Plant Mol. Biol. 17, 111–123.

[12] Ehrenhofer-Murray, A.E. (2004) Chromatin dynamics at DNA replication, transcription and repair. Eur. J. Biochem. 271, 2335–2349.

[13] Caiafa, P. and Zampieri, M. (2004) DNA methylation and chromatin structure: the puzzling CpG islands. J. Cell Biochem. 94, 257–265.

[14] Reik, W., Dean, W. and Walter, J. (2001) Epigenetic reprogramming in mammalian development. Science 293, 1089–1093.

[15] Li, E. (2002) Chromatin modification and epigenetic reprogramming in mammalian development. Nat. Rev. Genet. 3, 662–673.

[16] Hamet, P. and Tremblay, J. (2003) Genes of aging. Metabolism 52, 5–9.

[17] Ahuja, N. and Issa, J.P. (2000) Aging, methylation and cancer. Histol. Histopathol. 15, 835–842.

[18] Zhang, Z., Deng, C., Lu, Q., Richardson, B., Ahuja, N. and Issa, J.P. (2002) Age-dependent DNA methylation changes in the ITGAL (CD11a) promoter. Mech. Ageing Dev. 123, 1257–1268.

[19] Scarano, M.I., Strazzullo, M., Matarazzo, M.R. and D'Esposito, M. (2005) DNA methylation 40 years later: its role in human health and disease. J. Cell Physiol. 2004, 21–35.

[20] Jones, P.A., Ahuja, N. and Issa, J.P. (2002) DNA methylation and cancer. Oncogene 21, 5358–5360.

[21] Gaudet, F., Hodgson, J.G., Eden, A., Jackson-Grusby, L., Dausman, J., Gray, J.W., Leonhardt, H. and Jaenisch, R. (2003) Induction of tumors in mice by genomic hypomethylation. Science 300, 489–492.

[22] Eden, A., Gaudet, F., Waghmare, A. and Jaenisch, R. (2003) Chromosomal instability and tumors promoted by DNA hypomethylation. Science 300, 455.

[23] Issa, J.P. (2004) Opinion: CpG island methylator phenotype in cancer. Nat. Rev. Cancer 4, 988–993.

[24] Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A.R., Gutig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Muller, V., Otto, T., Scholz, C., Schulz, W.A., Seifert, H.H., Schwope, I., Ziebarth, H., Berlin, K., Piepenbrock, C. and Olek, A. (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res. 30, e21.

[25] Amoreira, C., Hindermann, W. and Grunau, C. (2003) An improved version of the DNA Methylation database (MethDB). Nucleic Acids Res. 31, 75–77.

[26] Joachims, T. (1999) in: Advances in Kernel Methods – Support Vector Learning (Smola, A., Schölkopf, B. and Burges, C., Eds.), MIT Press, Cambridge, MA; London, England.

[27] Zell, A. and Mamier, G. (1997) Stuttgart Neural Network Simulator, Version 4.2, University of Stuttgart, Stuttgart, Germany.

[28] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning representations by back-propagation errors. Nature 323, 533–536.

[29] Frank, E., Hall, K., Trigg, L., Holmes, G. and Witten, I.H. (2004) Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.

[30] Domingos, P. and Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130.

[31] Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression, John Wiley, New York.

[32] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.

[33] Dasarathy, B.V. (1991) Nearest neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press Tutorial.

[34] Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

[35] Joachims, T. (1999) in: (Smola, A., Schölkopf, B., Burges, C., Eds.), MIT Press, Cambridge, MA; London, England.

[36] Ritchot, N. and Roy, P.H. (1990) DNA methylation in Neisseria gonorrhoeae and other Neisseriae. Gene 86, 103–106.

[37] Shi, T., Liu, W.Z., Gao, F., Shi, G.Y and Xiao, S.D. (2005) Intranasal CpG-oligodeoxynucleotide is a potent adjuvant of vaccine against *Helicobacter pylori*, and T helper 1 type response and interferon-gamma correlate with the protection. Helicobacter 10, 71–79.

[38] Ward, J.J., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2003) Secondary structure prediction with support vector machines. Bioinformatics 19, 1650–1655.

[39] Bhasin, M. and Raghava, G.P.S. (2004) SVM based method for predicting HLADRB1*0401 binding peptides in an antigen sequence. Bioinformatics 20, 421–423.

[40] Brown, M.P.S., Grundy, W.N., Lion, D., Cristianini, N., Sugnet, C.W., Furey, T.S., AresJr, M. and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 97, 262–297.

[41] Costello, J.F., Fruhwald, M.C., Smiraglia, D.J., Rush, L.J., Robertson, G.P., Gao, X., Wright, F.A., Feramisco, J.D., Peltomaki, P., Lang, J.C., Schuller, D.E., Yu, L., Bloomfield, C.D., Caligiuri, M.A., Yates, A., Nishikawa, R., Su Huang, H., Petrelli, N.J., Zhang, X., O'Dorisio, M.S., Held, W.A., Cavenee, W.K. and Plass, C. (2000) Aberrant CpG-island methylation has non-random and tumor-type-specific patterns. Nat. Genet. 24, 132–138.

[42] Feltus, F.A., Lee, E.K., Costello, J.F., Plass, C. and Vertino, P.M. (2003) Predicting aberrant CpG island methylation. Proc. Natl. Acad. Sci. USA 100, 12253–12258.