Talanta 85 (2011) 1143–1147

Contents lists available at ScienceDirect

Talanta

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

Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition

Xuan Zhou[a,b], Zhanchao Li[a], Zong Dai[a], Xiaoyong Zou[a,*]

[a] School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, PR China
[b] School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou 510006, PR China

ARTICLE INFO

Article history:
Received 9 March 2011
Received in revised form 9 May 2011
Accepted 19 May 2011
Available online 27 May 2011

Keywords:
DNA methylation
Pseudo-trinucleotide composition
Support vector machine

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. The detections of DNA methylation have been determined mostly by experimental methods; however, these methods were time-consuming, expensive, and difficult to meet the requirements of modern large-scale sequencing technology. Accordingly, it is necessary to develop automatic and reliable prediction methods for DNA methylation.

In this study, the pseudo-trinucleotide composition was proposed, and a novel method was developed by support vector machine (SVM) with the pseudo-trinucleotide composition as input parameter to represent DNA sequence for DNA methylation prediction. The model was evaluated on two datasets, including a dataset of Rollins (dataset_1) and a dataset collected healthy human records from the MethDB database (dataset_2). For dataset_1, the Matthews correlation coefficient (*MCC*) and accuracy (*ACC*) by jackknife validation were 0.8051 and 0.6098, respectively. For dataset_2, the *MCC* and *ACC* were 0.8500 and 0.7203, respectively. The good prediction results reveal that the pseudo-trinucleotide composition is an effective representation method for DNA sequence and plays a very important role in the prediction of DNA function.

1. Introduction

In vertebrates, the methylation of cytosine is the most frequent endogenous modification of DNA, which consists of the addition of a methyl group to carbon-5 in the pyrimidine ring of cytosine generally in 5′-CpG-3′ dinucleotides, and is mediated by specific DNA-methyl transferases (DNMT) [1–3].

The methylation pattern of genes at CpG dinucleotide sites plays crucial roles in DNA function, such as DNA replication and repair [4], genetic imprinting [5], embryogenesis [6], gene transcription [7], and regulation of gene expression [7]. Genomic methylation patterns in non-dividing somatic differentiated cells are generally stable and heritable. However, if the methylation patterns undergo significant changes, the phenotype may be altered. For example, genome-wide changes in methylation patterns occur during developmental embryogenesis and in stem cell differentiation [8,9]; methylation patterns change in CpG islands of gene promoters during aging [10–12]; aberrant patterns of methylation have been found in various diseases [13], most notoriously in cancers [14]. Obviously, the evaluation of genomic DNA methylation status is of great significance.

In general, experimental methods of methylated CpG detection are laborious and time consuming. The experimental methods mainly include bisulfite conversion [15], methylation-sensitive restriction endonuclease cleavage assay [16], methylation-sensitive single nucleotide primer extension (Ms-SNuPE) [17], methylation-specific PCR (MSP) [18], etc. As a complementary technology for experimental detection, the computational prediction method can dig out the hidden important feature from data and provide important basis and research ideas for further understanding of the DNA methylation mechanism. Progresses have been made toward DNA methylation predicting. Bhasin et al. [19] developed an online predictive tool called “methylator”, which was a SVM-based method for the prediction of cytosine methylation in CpG dinucleotides by representing each nucleotide using conventional binary sparse encoding. Fang et al. [20] developed a classifier called “MethCGI” for predicting methylation status of CpG islands fragment using a SVM, and the nucleotide sequence contents as well as transcription factor binding sites (TFBSs) were used as features for the classification. Das et al. [21] developed a computational methylation pattern recognition classifier called “HDMFINDER”, which can be applied both to CpG islands and non-CpG island fragments by representing the DNA sequence

* Corresponding author. Tel.: +86 20 84114919; fax: +86 20 84112245.
E-mail addresses: ceszxy@mail.sysu.edu.cn, veego_z@hotmail.com (X. Zou).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.05.043

with specific sequence features such as GC content and using SVM for modeling. In these reported methods, it is crucial for DNA sequences to be parameterized by corresponding representation features for modeling.

Obviously, the representation methods of DNA sequence are the key for DNA methylation prediction. The existing representation features of DNA sequence mainly include nucleotide sequence contents such as GC content, CpG ratio, TpG content, di- and trinucleotide count, Alu coverage, transcription factor binding sites (TFBSs). These features can represent DNA sequence to a certain extent, but the sequence information will be more comprehensive if the features representing DNA sequences could consider sequence correlation factors. In general, more comprehensive information means better prediction.

Consequently, a novel representation method of DNA sequence – pseudo-trinucleotide composition was proposed as the input parameter for SVM to predict DNA methylation. In pseudo-trinucleotide composition, more sequence order effects were considered by introducing a series of sequence correlation factors with different tiers of correlation to improve the prediction.

2. Materials and methods

2.1. Dataset

Dataset_1: The dataset of Rollins et al. [22] is a large-scale description of the DNA methylation landscape of the human brain. It consists of the sequences of methylated and unmethylated DNA domains originally obtained by different sets of endonucleases. Methylated sequence libraries were created by digestion with the methylation-sensitive restriction endonucleases TaiI (ACGT), BstUI (CGCG), HhaI (GCGC), HpaII (CCGG) and AciI (CCGC and GCGG), unmethylated sequence libraries were created by digestion with McrBC (Rm^5C-N_{40-500}-Rm^5C). This dataset include thousands of methylated sequences and unmethylated sequences, each sequence is about 10–20 kilobases long. The dataset_1 can be downloaded freely from http://rulai.cshl.edu/HDMFinder/supplemental.htm. This website was built by Zhang Lab of Cold Spring Harbor Laboratory, which describes a computational pattern recognition method "HDMFINDER" [21] to predict the genomic DNA methylation profiles in the human adult brain. The algorithm computes the methylation propensity (methylated or unmethylated) for an 800 bp sequence region centered around a CpG dinucleotide based on specific sequence features within the region.

Dataset_2: We collected records of healthy humans in the MethDB database [23] and obtained 400 sequences. For each sequence, the actual methylation status was obtained according to 'm-score', which is defined as the proportion of m^5CpGs out of all CpGs in the region. If the m-score is less than 0.5, the corresponding sequence is defined as methylation-resistant; otherwise the sequence is regarded as methylation-prone. The dataset can be divided into 214 methylation-prone sequences and 186 methylation-resistant sequences. Each sequence is several hundred bases long, and mostly ranges from 100 to 600.

Methylation status predictions of DNA sequence include the predictions for CpG islands (CGIs) and non-CpG islands fragments (non-CGIs). However, the predictions for CGIs and non-CGI are confronted with two problems. Firstly, the predictions require a specific definition for CGI. Until now, the CGI criteria are not unique [24,25], for instance, the Gardiner-Garden criteria [24] is that a CGI should be no less than 200 bp with G+C content (%G+C) > 50% and (observed/expected) CpG ratio > 0.6; while the Takai–Jones CGI criteria [25] is that a CGI should be more than 500 bp with G+C content (%G+C) > 55% and (observed/expected) CpG ratio > 0.65. The "HDMFINDER" prediction tool [21] adopted the Takai–Jones CGI criteria, while the "MethCGI" [20] prediction tool adopted the Gardiner-Garden criteria. Secondly, the CGIs require the minimum length of DNA sequence, which implies the prediction is unsuitable for the short DNA sequences. In this work, considering the methylation predictions for CGIs and non-CGIs must meet some requirements, including the minimum of G+C content, the minimum of CpG ratio and the minimum length of DNA sequence, the methylation of whole DNA sequence was predicted without the classification of CGIs and non-CGIs.

2.2. Pseudo-trinucleotide composition

Since the amino acid composition can present protein amino acid sequence effectively [26,27], the nucleotide composition was proposed to present DNA base sequence. DNA sequences only consist of 4 nucleotides (A, C, G, T). If the DNA sequences were represented by single nucleotide composition, the characteristic parameters will be only 4, which is too few as input parameter, so the trinucleotide instead of single nucleotide composition was proposed. Although the trinucleotide composition includes some order information among the adjacent bases, it ignores the sequence order effects of whole DNA sequence. Therefore, the pseudo-trinucleotide composition was proposed for improving the methylation prediction accuracy.

The principle of pseudo-trinucleotide composition is introduced as follows according to pseudo-amino acid composition [28]. Suppose a gene fragment P with a sequence of L nucleotides:

$$P = R_1R_2R_3R_4R_5R_6R_7 \ldots R_L$$

where R_1 represents the nucleotide at chain position 1, R_2 the nucleotide at chain position 2, and so forth. Therefore, $R_1R_2R_3$ is the first corresponding trinucleotide, $R_2R_3R_4$ the second corresponding trinucleotide, and so forth. Since DNA sequences consist of 4 nucleotides (A, C, G, T), the array modes of three nucleotide have 64 ($4 \times 4 \times 4$) possibilities. Consequently, a gene fragment can be represented by the trinucleotide composition, a vector of 64 dimensions as S:

$$S = [f_1, f_2, f_3, f_4, f_5, f_6, \ldots, f_{64}] \quad (1)$$

where f_1 represents the occurrence frequency of the trinucleotide AAA in the gene fragment, f_2 the occurrence frequency of AAC, f_3 the occurrence frequency of AAG, and so forth.

The pseudo-trinucleotide composition can be expressed by a vector of $64 + \lambda$ dimensions as X. The first 64 components reflect the effect of the trinucleotide composition, whereas the components from $64 + 1$ to $64 + \lambda$ reflect the effect of sequence order.

$$X = [P_1, \ldots, P_{64}, P_{64+1}, \ldots, P_{64+\lambda}]^T \quad (\lambda < L) \quad (2)$$

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{64} f_i + w\sum_{k=1}^{\lambda} t_k}, & 1 \le u \le 64 \\ \dfrac{wt_{u-64}}{\sum_{i=1}^{64} f_i + w\sum_{k=1}^{\lambda} t_k}, & 64+1 \le u \le 64+\lambda \end{cases} \quad (3)$$

where f_i is the ith occurrence frequency of the 64 trinucleotides in the DNA sequence, t_k is the k-tier sequence correlation factor computed according to Eqs. (4) and (5), and w is the weight factor for the sequence order effect.

$$t_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k} \quad (k < L) \quad (4)$$

$$J_{i,i+k} = [H(R_{i+k}) - H(R_i)]^2 \quad (5)$$

In Eq. (4), t_1 is called the first-tier correlation factor that reflects the sequence order correlation between all the first most contiguous bases along DNA sequence, t_2 is the second-tier correlation factor that reflects the sequence order correlation between all the second most contiguous bases, and so forth. In Eq. (5), $H(R_i)$ is the physicochemical property value of base R_i. The values of physicochemical property were all subjected to a standard conversion as described by Eq. (6).

$$H(R_i) = \frac{H^0(R_i) - ave(H^0)}{SD(H^0)} \quad (6)$$

where *ave* represents the average value of the physicochemical property values of four bases, *SD* represents standard deviation. Here, the solvation free energy in water [29], which is the energy released when ions in crystal lattices associate with molecules in a solvent, was adopted as the physicochemical property.

The pseudo-trinucleotide composition contains more sequence order effects not only than the 64-D trinucleotide composition, as reflected by a series of sequence correlation factors with different tiers of correlation. These factors are defined by a correlation function that allows users to introduce any other biochemical quantities to obtain the optimal results for various cases concerned.

The conventional trinucleotide composition contains 64 discrete numbers, each of which reflects the occurrence frequency of one of the 64 trinucleotides in a DNA sequence. For the pseudo-trinucleotide composition, however, there are some other elements in addition to the 64 components. It is through these additional discrete numbers that the sequence order effect of a DNA sequence is approximately reflected and improvements are made. The physicochemical properties of DNA bases are useful for understanding the nucleic acid interactions and the stability of nucleic acid structures. In this work, the solvation free energy was introduced into the pseudo-trinucleotide composition, and it is anticipated that the sequence order effect of DNA sequence can be well represented.

2.3. Support vector machine (SVM)

In this study, SVM was chosen as the modeling method. SVM as a novel type of learning machine is gaining rapid popularity due to its remarkable generalization performance. The basic idea of SVM is to map the original data into a higher-dimensional feature space via a kernel function and then to perform classification in this space by constructing an optimal separating hyperplane. The SVM was initially developed for binary classification problems, and now SVM can also be utilized to solve nonlinear regression estimation by the introduction of ε-insensitive loss function. A detailed depiction to the theory of SVM for classification and regression can be referred to the literatures [30,31].

The public available LIBSVM software [32] can be downloaded freely from http://www.csie.ntu.edu.tw/~cjlin/libsvm. The radial basis function (RBF) was selected as the kernel function because of its effectiveness and speed in training process. The kernel parameters, including the penalty constant *C* and the parameters in kernel function (width parameter γ of radial basis function), need to be optimized in the modeling process.

2.4. Evaluation of the predictive performance

Prediction performance was determined by measuring threshold-dependent parameters sensitivity (*SE*), specificity (*SP*), accuracy (*ACC*) and Matthews correlation coefficient (*MCC*). *SE*, *SP*, *ACC* and *MCC* parameters were calculated by Eqs. (7)–(10), respectively.

$$SE = \frac{TP}{TP + FN} \quad (7)$$

$$SP = \frac{TN}{TN + FP} \quad (8)$$

$$ACC = TP + \frac{TN}{TP + FN + TN + FP} \quad (9)$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (10)$$

where *TP* are true positive (methylated site predicted as methylated); *FN* are false negative (methylated site predicted as non-methylated); *TN* are true negative (non-methylated site predicted as non-methylated) and *FP* are false positive (non-methylated site predicted as methylated).

3. Results and discussion

3.1. Selection of kernel function

Three common kernel (linear, polynomial and RBF) functions of SVM have been chosen for investigation on the datasets. The RBF kernel function was finally selected for the development of the current method. There are two parameters associated with SVM training. One is regularization of the cost parameter *C*, the other is kernel parameter γ, which determines the RBF width. The kernel parameters were optimized by 5-fold cross validation with the optimization of λ and *w* of the pseudo-trinucleotide composition simultaneously.

3.2. Optimization of pseudo-trinucleotide composition

In the construction process of pseudo-trinucleotide composition, the choice of the base's physicochemical properties has an impact on the prediction accuracy. In this work, the physicochemical properties of base such as the molecular weight, hydrophilicity, base stacking interaction, π-electron resonance energy [33,34], solvation free energy [29], flexibility [35], scores of generalized base properties (SGBP) [36] were investigated and the result from the solvation free energy was the optimum. Therefore, the solvation free energy which was involved in protein–nucleic acid interactions and the stability of nucleic acid tertiary structures [29], was selected for the construction of the pseudo-trinucleotide composition to predict DNA methylation.

By considering the sequence order effects of DNA sequence, the pseudo-trinucleotide composition can represent DNA sequence better than trinucleotide composition and achieve better prediction results. In pseudo-trinucleotide composition, there are two parameters of λ and *w* needed to be optimized. The *w* is in the range from 0 to 1, while the λ should be no longer than the length of the shortest DNA sequence in the datasets. Generally, the greater the λ, the more order information included in the pseudo-trinucleotide composition, however, the too large λ means redundant information and may lead to poorer prediction performance and longer training time. Therefore, the choice of suitable λ and *w* is crucial to the prediction performance. The 5-fold cross validation was used to optimize the λ and *w* by the prediction accuracy. And the optimizations of λ and *w* for dataset_1 and dataset_2 are shown in Figs. 1 and 2, respectively.

For dataset_1, when the value of λ is 0, namely the pseudo-trinucleotide composition is actually the trinucleotide composition, the prediction accuracy is 0.6502. When the value of λ is 5, as a result of the introduction of order information of DNA sequence, the

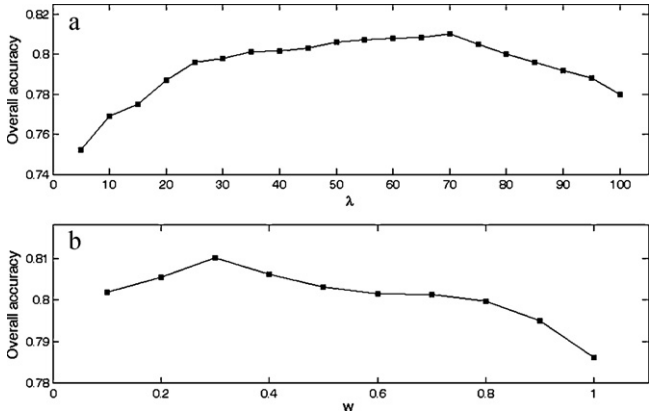

Fig. 1. The effect of λ(a) and *w*(b) on the prediction accuracy by 5-fold cross validation for dataset_1.

prediction accuracy increased to 0.7521. With the increase of λ, the prediction accuracy increased gradually. When the value of λ is 70, the prediction accuracy reached the maximum of 0.8101. Then the optimized pseudo-trinucleotide composition is a vector of 64 + 70 dimensions. As shown in Fig. 1(b), we can see that the prediction accuracy varied with *w*, and the prediction accuracy reached the maximum when *w* is 0.3. From above, the optimized λ and *w* were optimized as 70 and 0.3, respectively.

The optimization process for dataset_2 is illustrated in Fig. 2. As shown in Fig. 2(a), when the value of λ is 0, the prediction accuracy is 0.8150. Subsequently, the prediction accuracy improved with the increase of λ and reaches the maximum (0.8450) at the λ of 50. In Fig. 2(b), the prediction accuracy achieved the maximum when *w* is 0.3. That is to say, the optimized λ and *w* for dataset_2 is 50 and 0.3, respectively. Then the optimized pseudo-trinucleotide composition for dataset_2 is a vector of 64 + 50 dimensions.

In the definition of pseudo-trinucleotide composition, the pseudo-trinucleotide composition is a vector of 64 + λ dimensions. The first 64 components reflect the effect of the trinucleotide composition, whereas the components from 64 + 1 to 64 + λ reflect the effect of sequence order. The improvement of prediction accuracy of pseudo-trinucleotide composition is mainly embodied in that the sequence order effect was taken into account. The longer the length of the DNA sequence is, the more important the sequence order effect should be considered, therefore the optimized λ of different datasets may have some differences.

In this paper, dataset_1 and dataset_2 lead to different values for the optimal λ, which can be attributed to the different base lengths of these two datasets. In dataset_1, each sequence is about

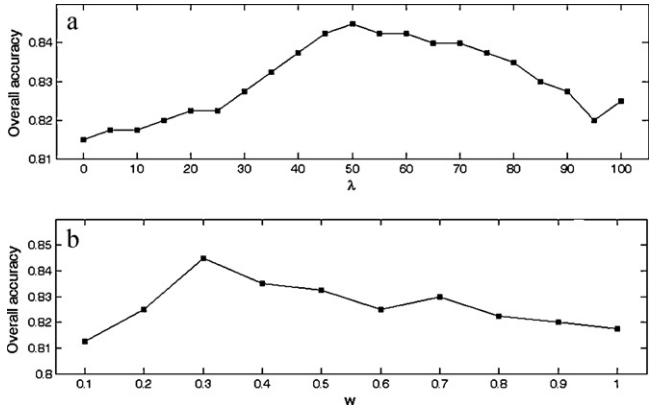

Fig. 2. The effect of λ(a) and *w*(b) on the prediction accuracy by 5-fold cross validation for dataset_2.

Table 1
Prediction accuracies of methylation for the whole DNA sequence.

Dataset	*SE*	*SP*	*ACC*	*MCC*
Dataset_1				
5-cross	0.7173	0.8874	0.8101	0.6181
Jackknife	0.7010	0.8920	0.8051	0.6098
Dataset_2				
5-cross	0.8925	0.7903	0.8450	0.7117
Jackknife	0.8972	0.7957	0.8500	0.7203

10–20 kilobases long, while in dataset_2, each sequence is several hundred bases long. Therefore, it is not difficult to explain why the dataset_1 need larger value of λ (more sequence order effect) than dataset_2. And actually, when the λ was selected between 50 and 70, the prediction accuracies of the two datasets in our paper have minor changes (0.8062–0.8101 for dataset_1 and 0.845–0.840 for dataset_2) with the variation of λ and the lost is little if the nonoptimal value in this range is selected, which means that a certain range of λ may be a consideration as the generic parameter.

3.3. Predictive performance

Since the jackknife test is deemed as one of the most rigorous and objective method in statistics, the optimized pseudo-trinucleotide composition was utilized to perform jackknife test for two datasets. The results are listed in Table 1. For dataset_1 and dataset_2, the values of *ACC* obtained from jackknife test are 0.8051 and 0.8500, respectively.

The proposed prediction methods mentioned above are for the DNA sequences without fragmenting. That is to say, whether the DNA sequences are CpG islands or not, the pseudo-trinucleotide composition can predict methylation status of human DNA sequence effectively. It means that the proposed method has no request for the definition of CpG islands and the length of DNA sequence. It has a special meaning to dataset_2, in which the length of DNA sequence is too short to be fragmented for CpG islands.

3.4. Comparison with other methods

The proposed method can also predict the methylation statuses of CpG islands effectively. To further validate our method, we compare it with the "HDMFINDER". In "HDMFINDER", each window centered around a CpG in the DNA sequences of the dataset of Rollins was extracted and tested whether it satisfies the Takai–Jones CGI criteria. Therefore, two datasets were composed, one for CGIs and another for non-CGIs. The SVM was utilized to predict their methylation status and two classifiers were obtained. When the window size is 800 bp, the optical prediction accuracy of CGIs and non-CGIs were 96.5% and 84%, respectively. For comparison, the methylation statuses of the CGIs and non-CGIs sample sets were predicted in our work by using the pseudo-trinucleotide composition as the features for SVM. The new prediction accuracies of CGIs and non-CGIs were respective 98.46% and 88.59%, 1.96% and 4.59% higher than the reported method. The results are also listed in Table 2.

Table 2
Prediction accuracies of methylation for CGIs and non-CGIs compared with "HDMFINDER".

	SE	*SP*	*ACC*	*MCC*
HDMFINDER (CGIs)	0.98	0.95	0.965	0.9325
Our method	0.9837	0.9854	0.9846	0.9692
HDMFINDER (non-CGIs)	0.81	0.87	0.84	0.6987
Our method	0.8676	0.9098	0.8859	0.7719

The existing algorithms for predicting DNA methylation were mainly based on the nucleotide sequence contents such as GC content, CpG ratio, di- and trinucleotide count, in which little sequence order effect was taken into account. To improve the prediction quality, it is necessary to incorporate such an effect. However, the number of possible patterns for DNA sequences is large, which has posed a formidable difficulty for realizing this goal. The pseudo-trinucleotide composition, a combination of a set of discrete sequence correlation factors and the 64 components of the trinucleotide composition, was proved to have the ability to deal with such a difficulty. It is anticipated that the concept of pseudo-trinucleotide composition and its mathematical framework and biochemical implication may have a series of impacts on the DNA methylation prediction and other areas of DNA functions as well.

4. Conclusion

In this paper, a novel DNA sequence representation method – pseudo-trinucleotide composition was proposed. The pseudo-trinucleotide composition was utilized to model SVM for the prediction of CpG status in humans DNA sequence, and the results indicated that the proposed method had the ability to achieve good prediction accuracy. It can be anticipated that the DNA sequence representation method of pseudo-trinucleotide composition may hold a high potential to become a useful tool for predicting other DNA functions.

Acknowledgments

We gratefully acknowledge to the financial support by the National Natural Science Foundation of China (Nos. 20975117, 20805059), the Natural Science Foundation of Guangdong Province (10151027501000070), the Scientific Technology Project of Guangdong Province (No. 2010A040302001), the Scientific Research Foundation for the Returned Overseas Chinese Scholars of State Education Ministry, and the Ph.D. Programs Foundation of Ministry of Education of China (No. 20070558010), the Fundamental Research Funds for the Central Universities (10lgzd13).

References

[1] W. Doerfler, Annu. Rev. Biochem. 52 (1983) 93–124.
[2] A. Hermann, H. Gowher, A. Jeltsch, Cell. Mol. Life Sci. 61 (2004) 2571–2587.
[3] A. Bird, J. Boyes, Cell 70 (1992) 5–8.
[4] P.A. Jones, S.B. Baylin, Nat. Rev. Genet. 3 (2002) 415–428.
[5] F. Feng, H. Wang, L. Han, S. Wang, J. Am. Chem. Soc. 30 (2008) 11338–11343.
[6] L. Song, S.R. James, L. Kazim, A.R. Karpf, Anal. Chem. 77 (2005) 504–510.
[7] A.P. Wolffe, M.A. Matzke, Science 286 (1999) 481–486.
[8] W. Reik, W. Dean, J. Walter, Science 293 (2001) 1089–1093.
[9] E. Li, Nat. Rev. Genet. 3 (2002) 662–673.
[10] P. Hamet, J. Tremblay, Metabolism 52 (2003) 5–9.
[11] N. Ahuja, J.P. Issa, Histol. Histopathol. 15 (2000) 835–842.
[12] Z. Zhang, C. Deng, Q. Lu, B. Richardson, N. Ahuja, J.P. Issa, Mech. Ageing Dev. 123 (2002) 1257–1268.
[13] M.I. Scarano, M. Strazzullo, M.R. Matarazzo, M. D_Esposito, J. Cell. Physiol. 204 (2005) 21–35.
[14] P.A. Jones, N. Ahuja, J.P. Issa, Oncogene 21 (2002) 5358–5360.
[15] S.J. Clark, A. Statham, C. Stirzaker, P.L. Molloy, M. Frommer, Nat. Protoc. 1 (2006) 2353–2364.
[16] A. Okamoto, K. Tanabe, I. Saito, J. Am. Chem. Soc. 124 (2002) 10262–10263.
[17] M.L. Gonzalgo, P.A. Jones, Nucleic Acids Res. 25 (1997) 2529–2531.
[18] J.G. Herman, J.R. Graff, S. Myohanen, B.D. Nelkin, S.B. Baylin, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 9821–9826.
[19] M. Bhasin, H. Zhang, E.L. Reinherza, P.A. Rechea, FEBS Lett. 579 (2005) 4302–4308.
[20] F. Fang, S.C. Fan, X.G. Zhang, M.Q. Zhang, Bioinformatics 22 (2006) 2204–2209.
[21] R. Das, N. Dimitrova, Z.Y. Xuan, R.A. Rollins, F. Haghighi, J.R. Edwards, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 10713–10716.
[22] R.A. Rollins, F. Haghighi, J.R. Edwards, R. Das, M.Q. Zhang, J. Ju, T.H. Bestor, Genome Res. 16 (2006) 157–163.
[23] C. Grunau, E. Renault, A. Rosenthal, G. Roizes, Nucleic Acids Res. 29 (2001) 270–274.
[24] G.M. Gardiner, M. Frommer, J. Mol. Biol. 196 (1987) 261–282.
[25] D. Takai, P.A. Jones, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 3740–3745.
[26] K.C. Chou, Proteins: Struct. Funct. Genet. 21 (1995) 319–344.
[27] K.C. Chou, D.W. Elrod, J. Proteome Res. 2 (2003) 183–190.
[28] K.C. Chou, Proteins 43 (2001) 246–255.
[29] M. Monajjemi, S. Ketabi, Z.M. Hashemian, A. Amiri, Biochemistry (Moscow) 71 (2006) S1–S8.
[30] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[31] N. Cristianini, T.J. Shawe, An Introduction to Support Vector Machines and Other 5kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.
[32] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/*cjlin/libsvm.
[33] B. Pullman, A. Pullman, Quantum Biochemistry, Wiley Interscience, New York, NY, 1963.
[34] O. Gotoh, Y. Takashlra, Biopolymers 20 (1981) 1033–1042.
[35] O.V. Shishkina, J. Sponerb, P. Hobzab, J. Mol. Struct. 477 (1999) 15–21.
[36] G.Z. Liang, Z.L. Li, J. Mol. Graph. Model. 26 (2007) 269–281.