Analytical Biochemistry 474 (2015) 69–77

Contents lists available at ScienceDirect

Analytical Biochemistry

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

iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition

Zi Liu [a], Xuan Xiao [a,b,*], Wang-Ren Qiu [a,*], Kuo-Chen Chou [b,c]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China
[b] Gordon Life Science Institute, Boston, MA 02478, USA
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 October 2014
Received in revised form 5 December 2014
Accepted 8 December 2014
Available online 14 January 2015

Keywords:
DNA methylation
3 → 1 Codon conversion
Pseudo amino acid components
Neighborhood cleaning rule
Synthetic minority oversampling technique
Target–jackknife cross-validation

ABSTRACT

Predominantly occurring on cytosine, DNA methylation is a process by which cells can modify their DNAs to change the expression of gene products. It plays very important roles in life development but also in forming nearly all types of cancer. Therefore, knowledge of DNA methylation sites is significant for both basic research and drug development. Given an uncharacterized DNA sequence containing many cytosine residues, which one can be methylated and which one cannot? With the avalanche of DNA sequences generated during the postgenomic age, it is highly desired to develop computational methods for accurately identifying the methylation sites in DNA. Using the trinucleotide composition, pseudo amino acid components, and a dataset-optimizing technique, we have developed a new predictor called "iDNA-Methyl" that has achieved remarkably higher success rates in identifying the DNA methylation sites than the existing predictors. A user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/iDNA-Methyl, where users can easily get their desired results. We anticipate that the web-server predictor will become a very useful high-throughput tool for basic research and drug development and that the novel approach and technique can also be used to investigate many other DNA-related problems and genome analysis.

Being a biochemical process where a methyl group is added to the cytosine residue, DNA methylation is involved in the regulation of many genes, plays an important role for epigenetic gene regulation in both life development and disease formation, and hence is considered a major epigenetic mark responsible for silencing of cell fate regulators [1]. In mammals, DNA methylation helps to regulate gene expression, genome imprinting, and X-chromosome inactivation [2].

Predominantly occurring on cytosine within a CG dinucleotide, DNA methylation is a covalent modification of DNA catalyzed by DNA methyltransferase (DNMT)[1] enzyme (Fig. 1). The DNA methylation sites are occupied by various proteins, including methyl-CpG binding domain (MBD) proteins that recruit a variety of histone deacetylase (HDAC) complexes and chromatin remodeling factors, leading to chromatin compaction and, consequently, to transcriptional repression. By either impeding the binding of transcriptional proteins to the gene [3] or bonding to the MBD [4], DNA methylation may affect the transcription of genes.

DNA methylation has long been suspected to play a role in tumorigenesis and cancer progression in various tissue types [5]. Actually, it plays a crucial role in developing nearly all types of cancer [6]. It is also important in stratifying patients for disease subclassification and personalized medicine, particularly in identifying biomarkers for improved therapeutic individualization [7]. Therefore, knowledge of DNA methylation sites is vitally important for both basic biomedicine research and practical drug development.

Cokus and coworkers [8] reported that there are some special sequence patterns for DNA methylation in the *Arabidopsis* genome. Using deep sequencing analysis, Kim and coworkers [9] revealed that there exist distinct patterns of DNA methylation in prostate cancer, indicating that the sites of DNA methylation are correlated with their sequential environments. In view of these findings, it is not only possible but also important to identify the methylation sites based on the sequence information alone. In particular, with

* Corresponding authors at: Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China (X. Xiao).

E-mail addresses: xxiao@gordonlifescience.org (X. Xiao), qiuone@163.com (W.-R. Qiu).

[1] *Abbreviations used:* DNMT, DNA methyltransferase; MBD, methyl-CpG binding domain; HDAC, histone deacetylase; SVM, support vector machine; KNN, *K*-nearest neighbor; AAC, amino acid composition; NAC, nucleic acid composition; PseAAC, pseudo amino acid composition; SMOTE, synthetic minority over-sampling technique; Acc, overall accuracy; MCC, Mathew's correlation coefficient; Sn, sensitivity; Sp, specificity; ROC, receiver operating characteristic.

http://dx.doi.org/10.1016/j.ab.2014.12.009
0003-2697/© 2014 Elsevier Inc. All rights reserved.

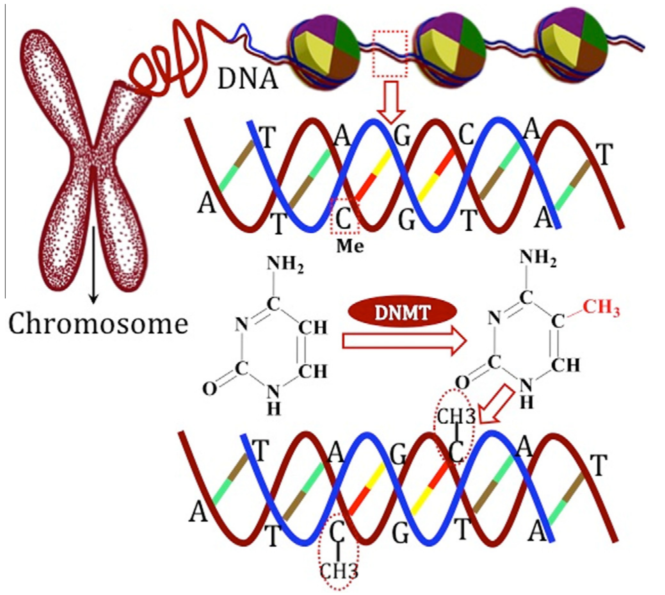


Fig. 1. A schematic drawing showing the process of DNA methylation. Catalyzed by DNA methyltransferase (DNMT), a methylation group is binding to the base cytosine (C) via a covalent bond.

the avalanche of DNA sequences generated during the postgenomic age, it is highly desired to develop computational methods for rapidly and effectively identifying the sites of DNA methylation.

In fact, during the past decade or so, some efforts have been made to use the computational approach to identify the DNA methylation sites. For instance, based on the support vector machine (SVM), Bhasin and coworkers [10] proposed a method called "Methylator" to predict the methylated CpGs in DNA sequences. Subsequently, Fang and coworkers [11] developed a method called "MethCGI" for predicting the methylation status of CpG islands in the human brain.

Although the aforementioned two predictors [10,11] did play an important role in stimulating the development of this area, they have the following shortcomings or limitations. First, in constructing the benchmark datasets for training and testing them, no clear cutoff procedure was imposed to remove the redundancy samples; accordingly, the two predictors could not avoid homology bias, and hence the success rates reported in Refs. [10,11] might be overestimated. Second, no sequence order information or effects were taken into account, and hence their prediction power might be limited. Third, no web-server whatsoever has been established for either the predictor Methylator [10] or the predictor MethCGI [11], which has significantly reduced the practical application value; particularly for most experimental scientists working in the fields of biomedicine and drug development, the impact of user-friendly web-servers is remarkable [12].

The current study was driven by the motivation to develop a new predictor in this regard by addressing the three drawbacks mentioned above.

Materials and methods

As shown by a series of recent publications [13–22] in response to the call from a comprehensive review [23] to develop and present a really useful statistical predictor for a biological system, one should make the following procedures crystal clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor, (ii) how to formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction, (iv) how to properly perform the cross-validation tests to objectively evaluate the anticipated accuracy of the predictor, and (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we address these five procedures one by one.

Benchmark datasets

For facilitating description later, we use the following scheme to represent a DNA sample:

$$\mathbf{D}_{\xi}(\mathbb{C}) = \mathrm{N}_{-\xi}\mathrm{N}_{-(\xi-1)}\cdots\mathrm{N}_{-2}\mathrm{N}_{-1}\mathbb{C}\mathrm{N}_{+1}\mathrm{N}_{+2}\cdots\mathrm{N}_{+(\xi-1)}\mathrm{N}_{+\xi} \tag{1}$$

where the center ($\mathbb{C}$) represents cytosine, the subscript ξ is an integer, $\mathrm{N}_{-\xi}$ represents the ξ-th upstream nucleotide from the center, N_{ξ} represents the ξ-th downstream nucleotide, and so forth (Fig. 2). The $(2\xi+1)$-tuple DNA sample $\mathbf{D}_{\xi}(\mathbb{C})$ can be further classified into the following categories:

$$\mathbf{D}_{\xi}(\mathbb{C}) \in \begin{cases} \mathbf{D}_{\xi}^{+}(\mathbb{C}), & \text{if its center is a methylation site} \\ \mathbf{D}_{\xi}^{-}(\mathbb{C}), & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{D}_{\xi}^{+}(\mathbb{C})$ represents a true methylation segment, $\mathbf{D}_{\xi}^{-}(\mathbb{C})$ represents a false methylation segment, and $\in$ represents "a member of" in the set theory.

As pointed out by a comprehensive review [24], there is no need to separate a benchmark dataset into a training dataset and a testing dataset if the predictor to be developed will be tested by the jackknife test or subsampling (*K*-fold) cross-validation test. Thus, the benchmark dataset for the current study can be formulated as:

$$\mathbb{S}_{\xi} = \mathbb{S}_{\xi}^{+} \cup \mathbb{S}_{\xi}^{-}, \tag{3}$$

where $\mathbb{S}_{\xi}^{+}$ contains only the samples of $\mathbf{D}_{\xi}^{+}(\mathbb{C})$ (i.e., the methylation DNA segments), $\mathbb{S}_{\xi}^{-}$ contains only the samples of $\mathbf{D}_{\xi}^{-}(\mathbb{C})$ (i.e., the non-methylation DNA segments; cf. Eq. (2)), and $\cup$ represents the symbol for union in the set theory.

Because the length of the peptide $\mathbf{D}_{\xi}(\mathbb{C})$ is $2\xi+1$ (cf. Eq. (1)), the benchmark dataset with different values of ξ will contain DNA segments of different numbers of nucleotides, as formulated by:

$$\mathbb{S}_{\xi} \text{ contains the segments of } \begin{cases} 37 \text{ nucleotides}, & \text{when } \xi = 18 \\ 39 \text{ nucleotides}, & \text{when } \xi = 19 \\ 41 \text{ nucleotides}, & \text{when } \xi = 20 \\ 43 \text{ nucleotides}, & \text{when } \xi = 21 \\ \vdots & \vdots \end{cases} \tag{4}$$

The detailed procedures to construct $\mathbb{S}_{\xi}$ are as follows. First, sliding a window of $(2\xi+1)$ nucleotides (Fig. 2) along each of the DNA sequences taken from MethDB [25] (http://www.methdb.de), a public database for DNA methylation, only those DNA segments with C (cytosine) at the center (i.e., the potential methylation site-containing segments) were collected. Second, if the upstream or downstream in a DNA was less than ξ, the lacking nucleotide was filled with the same nucleotide of its closest neighbor. Third, the DNA segment samples obtained in this way were put into the positive subset $\mathbb{S}_{\xi}^{+}$ if their centers had been experimentally annotated as the methylation sites; otherwise, they were put into the negative subset $\mathbb{S}_{\xi}^{-}$. Fourth, using CD-HIT software [26], the aforementioned samples were further subject to a screening procedure to winnow those that were identical to any other in a same subset. Fifth, excluded from the benchmark dataset were also those that were self-conflict, namely occurring in both methylation subset $\mathbb{S}_{\xi}^{+}$ and non-methylation subset $\mathbb{S}_{\xi}^{-}$.

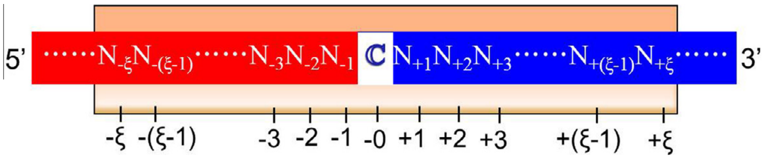

5'
······N-ξN-(ξ-1)······N-3N-2N-1
C
N+1N+2N+3······N+(ξ-1)N+ξ······
3'
-ξ
-(ξ-1)
-3
-2
-1
-0
+1
+2
+3
+(ξ-1)
+ξ

Fig.2. An illustration showing the flexible window $[-\xi, +\xi]$ sliding along a DNA sequence. Adapted from Chou [77] with permission. See the text for further explanation.

By following the aforementioned five steps and using $\xi = 18, 19, 20$, and 21 for the width of the sliding window, we obtained four benchmark datasets: $\mathbb{S}_{\xi=18}, \mathbb{S}_{\xi=19}, \mathbb{S}_{\xi=20}$, and $\mathbb{S}_{\xi=21}$, respectively. However, it was observed via preliminary trials that when $\xi = 20$ (i.e., the DNA segments concerned were formed with $20 \times 2 + 1 = 41$ nucleotides (cf. Eq. (4)), the corresponding results were most promising. Accordingly, we choose $\mathbb{S}_{\xi=20}$ as the benchmark dataset for further investigation. Thus, Eq. (3) can be hereafter expressed as:

$$\mathbb{S}(2426) = \mathbb{S}^{+}(787) \cup \mathbb{S}^{-}(1639) \tag{5}$$

where $\mathbb{S} = \mathbb{S}_{20} = \mathbb{S}(2426)$, which contains 2426 DNA segments, of which 787 samples are of methylation belonging to the positive dataset $\mathbb{S}^{+} = \mathbb{S}_{20}^{+} = \mathbb{S}(787)$, whereas 1639 are of non-methylation belonging to the negative dataset $\mathbb{S}^{-} = \mathbb{S}_{20}^{-} = \mathbb{S}(1639)$. The detailed sequences of the 787 + 1639 = 2426 DNA segments and their positions in the original DNA sequences are given in Supporting Information S1 of the online supplementary material.

Representation of DNA segment samples

The DNA samples in the current benchmark dataset can be generally expressed as:

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3 \cdots \mathrm{N}_i \cdots \mathrm{N}_{41}, \tag{6}$$

where N_i represents the nucleotide at the sequence position i $(1, 2, \cdots, 41)$. Based on the sequential model of Eq. (6), one could directly use BLAST [27] to perform statistical analysis. Although quite straightforward and simple, this kind of intuitive approach failed to work when a query sequence sample did not have significant similarity to any of the character-known sequences.

To cope with this problem, investigators could not help but resort to the discrete or vector model. Another reason for them to shift their efforts to develop various vector models is that samples formulated based on a vector model can be directly handled by all of the existing machine-learning algorithms such as the optimization approach [28], covariance discriminant (CD) [29,30], correlation coefficient method [31], neural network [32], SLLE (supervised locally linear embedding) algorithm [33], SVM [14,34], random forest [35], conditional random field [36], nearest neighbor (NN) [37], *K*-nearest neighbor (KNN) [38,39], optimized evidence-theoretic (OET)-KNN [40], fuzzy *K*-nearest neighbor [38], and multi-label (ML)-KNN algorithm [41]. Below, we elaborate how to develop an effective vector model for the current study and its rationale.

Just like using the amino acid composition (AAC) of a protein [42] to represent its sequence for statistical analysis, we can use the nucleic acid composition (NAC) to represent a DNA sample. Thus, a DNA sample in Eq. (6) can be expressed as:

$$\mathbf{D} = [\,f(\mathrm{A}) \quad f(\mathrm{C}) \quad f(\mathrm{G}) \quad f(\mathrm{T})\,]^{\mathbf{T}} \tag{7}$$

where $f(\mathrm{A}), f(\mathrm{C}), f(\mathrm{G})$, and $f(\mathrm{T})$ are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G), and thymine (T) in the DNA sequence, respectively, and the symbol $\mathbf{T}$ is the transpose operator. As we can see from Eq. (7), however, if using NAC to represent a DNA sample, all of its sequence order information would be completely lost.

Now, how can we formulate a DNA sequence with a vector yet considerably keep its sequence order information? Actually, a similar problem also occurred in dealing with the sequences of proteins and peptides, and hence it is one of the most fundamental problems in computational biology. One way to cope with such a problem is to represent the DNA sequence with the *k*-tuple nucleotide composition, that is:

$$\mathbf{D} = [\,f_1^{\text{K-tuple}} \quad f_2^{\text{K-tuple}} \quad \cdots \quad f_i^{\text{K-tuple}} \quad \cdots \quad f_{4^k}^{\text{K-tuple}}\,]^{\mathbf{T}} \tag{8}$$

where $f_i^{\text{K-tuple}}$ is the normalized occurrence frequency of the i-th k-tuple nucleotide in the DNA sequence. As we can see from Eq. (8), when $k > 3$ the number of the corresponding components will rapidly increase, causing the so-called "high-dimension disaster" problem [43]. To avoid the problem, here we used the 3-tuple nucleotide or trinucleotide composition (TNC) to formulate the DNA sample. Besides, doing so also has clearer biological meaning because, according to the $3 \rightarrow 1$ genetic rule (Fig. 3), a codon of three nucleotides in DNA defines an amino acid (Table 1) in protein. Thus, instead of Eq. (8), we have:

$$\begin{aligned} \mathbf{D} &= \left[f_1^{\text{3-tuple}} \quad f_2^{\text{3-tuple}} \quad f_3^{\text{3-tuple}} \quad f_4^{\text{3-tuple}} \quad \cdots \quad f_{64}^{\text{3-tuple}} \right]^{\mathbf{T}} \\ &= [\,f(\mathrm{AAA}) \quad f(\mathrm{AAC}) \quad f(\mathrm{AAG}) \quad f(\mathrm{AAT}) \quad \cdots \quad f(\mathrm{TTT})\,]^{\mathbf{T}} \end{aligned} \tag{9}$$

where $f_1^{\text{3-tuple}} = f(\mathrm{AAA})$ is the normalized occurrence frequency of AAA in the DNA sequence, $f_2^{\text{3-tuple}} = f(\mathrm{AAC})$ is that of AAC, $f_3^{\text{3-tuple}} = f(\mathrm{AAG})$, is that of AAG, and so forth. Using Eq. (9), however, we can incorporate only the local sequence order information between the most and second most contiguous nucleotides but not the global or long-range sequence order information.

To incorporate the long-range sequence order information of a DNA sample, let us adopt the concept of pseudo amino acid composition [44,45] or Chou's PseAAC [46,47] in protein. Since the concept of PseAAC was proposed in 2001 [44], it has been used in nearly all the fields of protein attribute predictions (see, e.g., Refs. [48–53] as well as a long list of publications cited in a recent review [54]. Because it has been widely used, recently three types of powerful open access software, called PseAAC-Builder [55], propy [56], and PseAAC-General [54], were established; the first two are for generating various modes of Chou's special PseAAC, whereas the third is for generating those of Chou's general PseAAC.

In a way parallel to the approach of using the PseAAC formulation for protein/peptide, a DNA sample can also be formulated with the PseTNC (pseudo trinucleotide composition), as given by [57]:

$$\mathbf{D} = [\,d_1 \quad d_2 \quad \cdots \quad d_{64} \quad d_{64+1} \quad \cdots \quad d_{64+\lambda}\,]^{\mathbf{T}} \tag{10}$$

where the first 64 components reflect the local or short-range sequence order effect and the next $\lambda = 8$ components reflect the global or long-range effects. Its detailed derivation, along with how to calculate the $64 + 8 = 72$ components in Eq. (10), was clearly elaborated in Ref. [57], and hence there is no need to repeat here.

Optimizing imbalanced benchmark datasets

The aforementioned benchmark dataset is very imbalanced; that is, the size of the negative subset $\mathbb{S}^{-}$ is more than two times

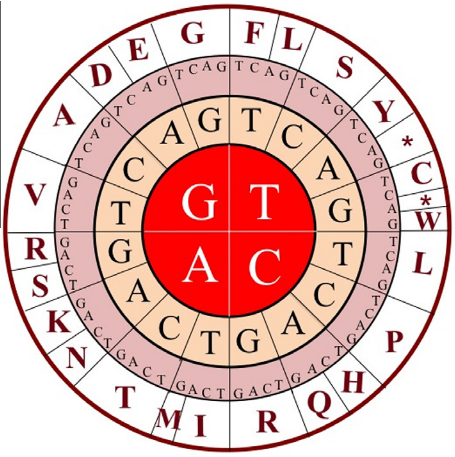


Fig.3. A graph showing how a DNA codon of three nucleotides is converted into an amino acid. The characters in the first three rings from the center represent four bases in DNA, whereas those in the fourth ring represent the single-letter codes of the 20 native amino acids in protein. The symbol * represents the "stop" sign.

the size of the positive subset $\mathbb{S}^+$. Although this might reflect the real world in which the non-methylation sites are always the majority compared with the methylation ones, a predictor trained with such a skewed benchmark dataset would have the consequence that many DNA methylation sites might be mispredicted as non-methylation ones [58]. Actually, what is really the most intriguing information for the drug development scientists is the information about the methylation sites. Therefore, it is important to find an effective approach to optimize the unbalanced benchmark dataset and minimize the consequence of this kind of misprediction [59].

To realize this, we used the NCR (neighborhood cleaning rule) [60] and the SMOTE (synthetic minority over-sampling technique) [61] treatments to optimize the aforementioned skewed benchmark datasets. The former is to remove some redundant negative samples from the negative subset so as to minimize its statistical noise, which can be likened to the sample screening procedure in computational proteomics (see, e.g., Ref. [62]). The latter is to add some hypothetical positive samples into the positive subset so as to enhance the ability in identifying the methylation sites, which can be likened to the seed propagation approach in Ref. [63] and the Monte Carlo sampling approach in Refs. [64,65] for expanding the positive subsets.

The detailed procedures for the NCR treatment [60] in the current study are as follows. First, for each of the samples in the benchmark dataset $\mathbb{S}$, find its three nearest neighbors. Second, if the sample concerned belongs to the negative subset $\mathbb{S}^-$ and at least two of its three nearest neighbors belong to the positive subset $\mathbb{S}^+$, remove the sample from the benchmark dataset. Third, however, if it belongs to the positive subset $\mathbb{S}^+$, remove those of its nearest neighbors that belong to the negative subset $\mathbb{S}^-$ from the benchmark dataset. After the aforementioned NCR treatment, 522 negative samples (cf. Supporting Information S2 in supplementary material) were removed from the negative subset.

Subsequently, to further optimize the benchmark dataset, the SMOTE approach [61] was adopted to create 330 hypothetical samples for the positive subsets by the linear interpolation scheme.

The above two treatments, one for removing 522 samples from the negative subset and the other for adding 330 hypothetical samples to the positive subset, along with the final outcome can be formulated as:

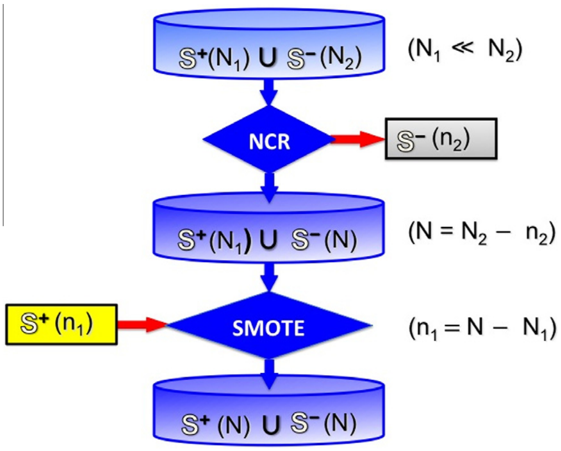


Fig.4. A flowchart showing the process of converting an imbalanced benchmark dataset into a balanced one by the NCR and SMOTE treatments. In the figure, N_1 and N_2 represent the numbers of samples in the original positive and negative subsets, respectively, n_2 represents the number of negative samples removed by the NCR treatment, and n_1 represents the number of positive hypothetical samples created by SMOTE and added to the positive subset to make the optimized benchmark dataset completely in balance between its two subsets. In the current case, we have $N_1 = 787$, $N_2 = 1639$, $n_1 = 330$, and $n_2 = 522$. See the relevant text for further explanation.

Table 1
Codon conversion from DNA trinucleotides to protein amino acids.

First base	Second base				Third base
	A	C	G	T	
A	Lys	Thr	Arg	Ile	A
	Asn	Thr	Ser	Ile	C
	Lys	Thr	Arg	Met	G
	Asn	Thr	Ser	Ile	T
C	Gln	Pro	Arg	Leu	A
	His	Pro	Arg	Leu	C
	Gln	Pro	Arg	Leu	G
	His	Pro	Arg	Leu	T
G	Glu	Ala	Gly	Val	A
	Asp	Ala	Gly	Val	C
	Glu	Ala	Gly	Val	G
	Asp	Ala	Gly	Val	T
T	Stop!	Ser	Stop!	Leu	A
	Tyr	Ser	Cys	Phe	C
	Stop!	Ser	Trp	Leu	G
	Tyr	Ser	Cys	Phe	T

$$\begin{cases} \mathbb{S}(2426) = \mathbb{S}(787) \cup \mathbb{S}(1639), & \text{from original Eq .(5)} \\ \mathbb{S}(1904) = \mathbb{S}(787) \cup \mathbb{S}(1117), & \text{after NCR treatment} \\ \mathbb{S}(2234) = \mathbb{S}(1117) \cup \mathbb{S}(1117), & \text{after SMOTE treatment} \end{cases} \tag{11}$$

Meanwhile, to provide an intuitive picture, a flowchart is given in Fig. 4 to illustrate the process of how to optimize an imbalance benchmark dataset.

For the reader's convenience, the 1117 positive samples and 1117 negative samples finally obtained by the optimization procedures are given in Supporting Information S3 of the supplementary material. Note that the aforementioned positive samples contain 330 hypothetical samples that were generated via the linear interpolation scheme in SMOTE and, hence, can be expressed only by their feature vectors as defined in Eq. (10) but not real sample codes as given in Supporting Information S1. Nevertheless, it would be perfectly fine to do so because the data directly used to train a predictor were actually the samples' feature vectors but not their original sequences. This is the key to optimize an imbalanced benchmark dataset with such a novel approach, and its rationale is further elucidated later.

SVM classifier

The SVM classifier or SVM algorithm [66] has been widely used in many areas of bioinformatics (see, e.g., Refs. [14,67–71]). The basic idea of SVM is to construct a separating hyper-plane to maximize the margin between the positive dataset and negative dataset. For a brief formulation of SVM and how it works, see Refs. [34,72]; for more details about SVM, see Ref. [73].

The software of SVM used in the current study was downloaded from the LIBSVM (library for SVMs) package [74,75], which contains two built-in parameters: c and γ. To maximize the performance, the two parameters were preliminarily optimized with the search function "SVMcgForClass" downloaded from http://www.matlabsky.com.

The DNA samples as formulated by Eq. (10) were used as inputs for the SVM classifier. Given a query vector sample, the classifier can quite accurately predict which class it belongs to after training by a relevant dataset, that is, clearly indicating whether it is a "methylation DNA segment" or "non-methylation DNA segment" (cf. Eq. (2)).

The predictor obtained via the aforementioned procedures is called "iDNA-Methyl," where "i" denotes "identify" and "DNA-Methyl" denotes "DNA methylation site."

Results and discussion

As pointed out in the beginning of Materials and Methods, one of the important procedures in developing a new predictor is how to properly and objectively evaluate its quality [23]. This actually comprises the following two aspects: (i) what metrics should be used to quantitatively measure the prediction accuracy and (ii) what kind of method should be adopted to do the test. Below, we address these problems.

A set of four metrics for measuring prediction quality

For the problem investigated in the current study, four indexes are often used in the literature: (i) overall accuracy (or Acc), (ii) Mathew's correlation coefficient (or MCC), (iii) sensitivity (or Sn), and (iv) specificity (or Sp) (see, e.g., Ref. [76]). However, the conventional formulations for the four metrics are not quite intuitive and easy to be understood by most experimental scientists, particularly the one for MCC. Actually, by using the symbols and derivation as used in Ref. [77] for studying signal peptides, the aforementioned four metrics can be formulated by a set of equations given below:

$$\begin{cases} \text{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \leqslant \text{Sn} \leqslant 1 \\ \text{Sp} = 1 - \frac{N_+^-}{N^-} & 0 \leqslant \text{Sp} \leqslant 1 \\ \text{Acc} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \leqslant \text{Acc} \leqslant 1 \\ \text{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-} \right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \leqslant \text{MCC} \leqslant 1 \end{cases} \tag{12}$$

where N^+ represents the total number of DNA methylation segments investigated, N_-^+ represents the number of true methylation segments incorrectly predicted as the non-methylation segments, N^- represents the total number of non-methylation segments investigated, and N_+^- represents the number of non-methylation segments incorrectly predicted as the methylation segments.

With Eq. (12) at hand, it is now crystal-clear to see the following. When $N_-^+ = 0$, meaning that none of the methylation segments is incorrectly predicted to be a non-methylation segment, we have the sensitivity $\text{Sn} = 1$. When $N_-^+ = N^+$, meaning that all of the methylation segments are incorrectly predicted to be non-methylation segments, we have the sensitivity $\text{Sn} = 0$. Likewise, when $N_+^- = 0$, meaning that none of the non-methylation segments was incorrectly predicted to be a methylation segment, we have the specificity $\text{Sp} = 1$. When $N_+^- = N^-$, meaning that all of the non-methylation segments were incorrectly predicted as methylation segments, we have the specificity $\text{Sp} = 0$. When $N_-^+ = N_+^- = 0$, meaning that none of methylation segments in the positive dataset and none of the non-methylation segments in the negative dataset was incorrectly predicted, we have the overall accuracy $\text{Acc} = 1$ and $\text{MCC} = 1$. When $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all of the methylation segments in the positive dataset and all of the non-methylation segments in the negative dataset were incorrectly predicted, we have the overall accuracy $\text{Acc} = 0$ and $\text{MCC} = -1$. When $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have $\text{Acc} = 0.5$ and $\text{MCC} = 0$, indicating no better than random prediction. As we can see from the above discussion based on Eq. (12), the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier to understand.

It should be pointed out, however, that the set of metrics as defined in Eq. (12) is valid only for the single-label systems. For the multi-label systems, whose emergence has become more frequent in system biology [78,79] and system medicine [80,81], a completely different set of metrics as defined in Ref. [41] is needed.

Jackknife and target–jackknife cross-validation

The following three cross-validation methods are often used to validate a statistical predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [82]. Of the three methods, however, the jackknife test is deemed the least arbitrary one that can always yield a unique outcome for a given benchmark dataset, as elucidated in Ref. [83] and demonstrated by Eqs. (28–30) therein. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., Refs. [33,51,53,84–87]). During the jackknife process, each of the samples in the benchmark dataset is singled out in turn and tested by the predictor trained by the remaining samples.

When carrying out the jackknife test on the optimized benchmark dataset of Eq. (11), however, some special consideration is needed. This is because it has contained 330 hypothetical positive

Table 2
A comparison of iDNA-Methyl with the existing predictors via the jackknife cross-validation on the same experiment-confirmed data.

Predictor	Acc (%)	MCC	Sn (%)	Sp (%)	Web-server
iDNA-Methyl[a]	77.49	54.71	61.25	90.33	Yes
Methylator[b]	71.35	33.27	51.72	80.78	Not working
MethCGI[c]	73.83	37.48	49.68	85.42	Not working

[a] Results obtained by the target–jackknife test for the current predictor on the experiment-confirmed data. See "Representation of DNA segment samples" section for further explanation about the target–jackknife test.

[b] Results obtained by the jackknife test for the predictor by Bhasin and coworkers [10] on the same experiment-confirmed data.

[c] Results obtained by the jackknife test for the predictor by Fang and coworkers [11] on the same experiment-confirmed data.

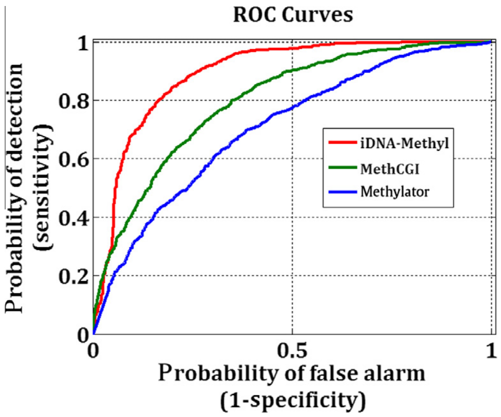


Fig.5. The ROC (receiver operating characteristic) curves showing the predictor's quality. The area under the ROC curve for iDNA-Methyl is obviously larger than those of its counterparts, indicating a clear improvement of the new predictor over the existing ones [93].

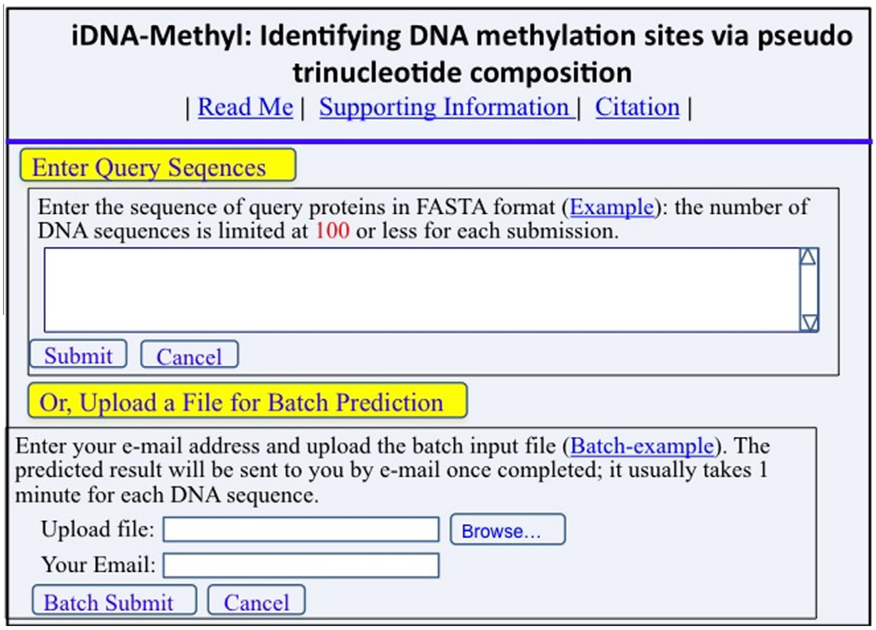
samples and excluded 505 experimental negative samples. Because the validation should be conducted strictly based on the experimental data only, a special jackknife test, the so-called target–jackknife test, has been introduced in this study. During the target–jackknife process, only the experiment-confirmed samples are in turn singled out as a target (or test sample) for cross-validation. Thus, when validating the predictor with the positive samples in $\mathbb{S}^{+}(1117)$, only $(1117 - 330) = 787$ experimental positive samples need to be singled out for cross-validation. When validating the predictor with the negative samples, however, counted are not only all 1117 samples in $\mathbb{S}^{-}(1117)$ but also the 522 experimental negative samples that have been removed from the optimized benchmark dataset by the NCR treatment. Doing so will make it absolutely fair in comparing the performance of the current predictor with the other existing methods, as elaborated in next section; that is, all of the predictors are tested using exactly the same experiment-confirmed samples.

Comparison with the existing predictors

The scores for the four metrics as defined in Eq. (12) achieved by the current iDNA-Methyl predictor via the target–jackknife tests are given in Table 2, where for facilitating comparison the corresponding scores by the existing predictors are also listed. From the table, we can observe the following. First, the score of overall accuracy (Acc) achieved by the current predictor iDNA-Methyl is remarkably higher than those achieved by the existing predictors [10,11]. Second, it is also true for the other three metrics, clearly indicating that the new predictor, in identifying DNA methylation sites, not only can yield higher prediction accuracy but also is more stable with higher sensitivity and specificity.

Because graphic approaches can provide useful intuitive insights (see, e.g., Refs. [88–92]), here we also provide a graphic comparison of the current predictor with their counterparts via the ROC (receiver operating characteristic) plot [93], as shown in Fig. 5. According to the ROC [93], the larger the area under the curve, the better the corresponding predictor. As we can see from the figure, the area under the ROC curve of the new predictor is remarkably greater than those of its counterparts, indicating a clear improvement of the new predictor in comparison with the existing ones.

It is instructive to point out that although the current predictor iDNA-Methyl was trained by the optimized benchmark dataset in which 522 experiment-confirmed negative samples were removed from the original negative subset to balance its size with the positive one, they were still counted in the target–jackknife cross-validation. On the other hand, although 330 hypothetical positive


iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition

| Read Me | Supporting Information | Citation |

Enter Query Seqences

Enter the sequence of query proteins in FASTA format (Example): the number of DNA sequences is limited at 100 or less for each submission.

Submit Cancel

Or, Upload a File for Batch Prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 1 minute for each DNA sequence.

Upload file: Browse...

Your Email:

Batch Submit Cancel

Fig.6. A semi-screenshot showing the top page of the iDNA-Methyl web-server at http://www.jci-bioinfo.cn/iDNA-Methyl.

samples were added into the optimal positive subset to make it completely in balance with the negative one, only the experimental samples were counted in calculating the metric scores. In other words, the experiment-confirmed samples counted during the cross-validation for calculating the metric scores of iDNA-Methyl, regardless of whether they are positive or negative samples, are exactly the same as those used to test the other predictors listed in Table 2.

Web-server and user guide

As shown in Table 2, none of the two existing predictors [10,11] has a working web-server. In contrast, a workable web-server was established for the new predictor, which is particularly important for those who are interested in using the iDNA-Methyl predictor but not its mathematical details. Below, we give a step-by-step guide on how to use the web-server to get the desired results.

Step 1

Open the web-server at http://www.jci-bioinfo.cn/iDNA-Methyl, and you will see the top page of the iDNA-Methyl predictor on your computer screen, as shown in Fig. 6. Click on the "Read Me" button to see a brief introduction about iDNA-Methyl and the caveat when using it.

Step 2

Either type or copy/paste the query DNA sequences into the input box at the center of Fig. 6. The input sequence should be in the FASTA format. For examples of sequences in FASTA format, click on the "Example" button right above the input box.

Step 3

Click on the "Submit" button to see the predicted result. For example, if you use the query DNA sequences in the "Example" window as the input, you will see the following shown on the screen of your computer: (1) DNA sequence 1 contains 160 C (cytosine) residues, of which only those at the sequence positions 17, 195, 209, and 223 are predicted to be the methylation sites, and all of the others are not. (2) DNA sequence 2 contains 294 C (cytosine) residues, of which only those at the sequence positions 378, 786, 797, 831, and 1017 are predicted to be the methylation sites, and all of the others are not. All of these results are fully consistent with the experimental observations.

Step 4

As shown on the lower panel of Fig. 6, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the "Browse..." button. To see the sample of batch input file, click on the "Batch-example" button. After clicking the "Batch Submit" button, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail."

Step 5

Click the "Supporting Information" button to download the benchmark dataset used to train and test the iDNA-Methyl predictor.

Step 6

Click on the "Citation" button to find the relevant articles that document the detailed development and algorithm of iDNA-Methyl.

Conclusion

We anticipate that the iDNA-Methyl predictor will become a useful high-throughput tool because (i) information of DNA methylation sites is important for both basic research and drug development and (ii) compared with the existing predictors in identifying the DNA methylation sites, it has remarkably higher success rates and a workable and publicly accessible web-server.

It has not escaped our notice that the approach of using the 3 → 1 codon conversion to incorporate the long-range or global sequence order information of DNA, as well as the technique of using NCR and SMOTE to optimize unbalanced datasets, can be effectively used in many other areas of genome analysis as well.

Acknowledgments

The authors thank the two anonymous reviewers, whose constructive comments were very helpful in strengthening the presentation of this study. This work was partially supported by the National Nature Science Foundation of China (31260273 and 61261027), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (20120BDH80023), the Natural Science Foundation of Jiangxi Province (20114BAB211013, 20122BAB211033, 20122BAB201044, and 20122BAB201020), the Department of Education of Jiangxi Province (GJJ12490), the LuoDi plan of the Department of Education of Jiangxi Province (KJLD12083), the Jiangxi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008), and the Graduated Innovation Fund of the Jingdezhen Ceramic Institute (JYC130).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ab.2014.12.009.

References

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

[2] J. Song, M. Teplova, S. Ishibe-Murakami, D.J. Patel, Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation, Science 335 (2012) 709–712.

[3] M.K. Choy, M. Movassagh, H.G. Goh, M.R. Bennett, T.A. Down, R.S. Foo, Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated, BMC Genomics 11 (2010) 519.

[4] B. Hendrich, C. Abbott, H. McQueen, D. Chambers, S. Cross, A. Bird, Genomic structure and chromosomal mapping of the murine and human Mbd1, Mbd2, Mbd3, and Mbd4 genes, Mamm. Genome 10 (1999) 906–912.

[5] Y. Kobayashi, D.M. Absher, Z.G. Gulzar, S.R. Young, J.K. McKenney, D.M. Peehl, J.D. Brooks, R.M. Myers, G. Sherlock, DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer, Genome Res. 21 (2011) 1017–1027.

[6] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, Nat. Genet. 33 (Suppl.) (2003) 245–254.

[7] E.C. Schwalbe, D. Williamson, J.C. Lindsey, D. Hamilton, S.L. Ryan, H. Megahed, M. Garami, P. Hauser, B. Dembowska-Baginska, D. Perek, DNA methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies, Acta Neuropathol. 125 (2013) 359–371.

[8] S.J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, S.E. Jacobsen, Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning, Nature 452 (2008) 215–219.

[9] J.H. Kim, S.M. Dhanasekaran, J.R. Prensner, X. Cao, D. Robinson, S. Kalyana-Sundaram, C. Huang, S. Shankar, X. Jing, M. Iyer, Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer, Genome Res. 21 (2011) 1028–1041.

[10] M. Bhasin, H. Zhang, E.L. Reinherz, P.A. Reche, Prediction of methylated CpGs in DNA sequences using a support vector machine, FEBS Lett. 579 (2005) 4302–4308.

[11] F. Fang, S. Fan, X. Zhang, M.Q. Zhang, Predicting methylation status of CpG islands in the human brain, Bioinformatics 22 (2006) 2204–2209.

[12] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. (in press). http://dx.doi.org/10.2174/1573406411666141229162834.
[13] W. Chen, P.M. Feng, H. Lin, IRSpot–PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[14] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.
[15] R. Xu, J. Zhou, B. Liu, Y.A. He, Q. Zou, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. (in press). doi: http://dx.doi.org/10.1080/07391102.
[16] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, IDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS ONE 9 (2014) e106691.
[17] W.R. Qiu, X. Xiao, W.Z. Lin, IMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, Biomed. Res. Int. 2014 (2014) 947416.
[18] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, INitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS ONE 9 (2014) e105018.
[19] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, INuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo *k*-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[20] W. Chen, P.M. Feng, E.Z. Deng, ITIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.
[21] H. Lin, E.Z. Deng, H. Ding, W. Chen, IPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo *k*-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[22] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, J. Biomol. Struct. Dyn. (in press). doi: http://dx.doi.org/10.1080/07391102.
[23] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition [50th anniversary year review], J. Theor. Biol. 273 (2011) 236–247.
[24] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction [review], Anal. Biochem. 370 (2007) 1–16.
[25] C. Amoreira, W. Hindermann, C. Grunau, An improved version of the DNA methylation database (MethDB), Nucleic Acids Res. 31 (2003) 75–77.
[26] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.
[27] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.
[28] C.T. Zhang, An optimization approach to predicting protein structural class from amino acid composition, Protein Sci. 1 (1992) 401–408.
[29] K.C. Chou, Prediction of G-protein-coupled receptor classes, J. Proteome Res. 4 (2005) 1413–1418.
[30] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins 50 (2003) 44–48.
[31] K.C. Chou, C.T. Zhang, A correlation coefficient method to predicting protein structural classes from amino acid compositions, Eur. J. Biochem. 207 (1992) 429–433.
[32] T.B. Thompson, C. Zheng, Neural network prediction of the HIV-1 protease cleavage sites, J. Theor. Biol. 177 (1995) 369–379.
[33] M. Wang, J. Yang, Z.J. Xu, SLLE for predicting membrane protein types, J. Theor. Biol. 232 (2005) 7–15.
[34] Y.D. Cai, G.P. Zhou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.
[35] K.K. Kandaswamy, T. Martinetz, S. Moller, P.N. Suganthan, S. Sridharan, G. Pugalenthi, AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties, J. Theor. Biol. 270 (2011) 56–62.
[36] Y. Xu, J. Ding, L.Y. Wu, ISNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS ONE 8 (2013) e55844.
[37] H.B. Shen, Using optimized evidence-theoretic *K*-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types, Biochem. Biophys. Res. Commun. 334 (2005) 288–292.
[38] X. Xiao, P. Wang, GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions, Mol. BioSyst. 7 (2011) 911–919.
[39] P. Wang, X. Xiao, NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features, PLoS ONE 6 (2011) e23505.
[40] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.
[41] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (2013) 1092–1100.
[42] K.C. Chou, A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space, Proteins 21 (1995) 319–344.
[43] T. Wang, J. Yang, H.B. Shen, Predicting membrane protein types by the LLDA algorithm, Protein Pept. Lett. 15 (2008) 915–921.
[44] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (Erratum: 2001, vol. 44, p. 60).
[45] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[46] S.X. Lin, J. Lapointe, Theoretical and experimental biology in one, J. Biomed. Sci. Eng. 6 (2013) 435–442.
[47] W.Z. Zhong, S.F. Zhou, Molecular science for drug development and biomedicine, Int. J. Mol. Sci. 15 (2014) 20072–20078.
[48] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.
[49] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, J. Theor. Biol. 257 (2009) 17–26.
[50] M.M. Beigi, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach, J. Struct. Funct. Genomics 12 (2011) 191–197.
[51] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. 341 (2014) 34–40.
[52] M. Khosravian, F.K. Faramarzi, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods, Protein Pept. Lett. 20 (2013) 180–186.
[53] H. Mohabatkar, M.M. Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach, Med. Chem. 9 (2013) 133–137.
[54] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.
[55] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2012) 117–119.
[56] D.S. Cao, Q.S. Xu, Y.Z. Liang, Propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.
[57] W.R. Qiu, X. Xiao, IRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.
[58] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recognit Artif Intell. 23 (2009) 687–719.
[59] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. (in press). doi: http://dx.doi.org/10.1080/07391102.
[60] J. Laurikkala, Improving identification of difficult small classes by balancing class distribution, in: AIME '01: Proceedings of the 8th Conference on AI in Medicine in Europe (pp. 63–66), Springer-Verlag, London, 2001.
[61] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357.
[62] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.
[63] C.T. Zhang, An analysis of protein folding type prediction by seed-propagated sampling and jackknife test, J. Protein Chem. 14 (1995) 583–593.
[64] C.T. Zhang, Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition, Biophys. J. 63 (1992) 1523–1529.
[65] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, J. Biol. Chem. 268 (1993) 16938–16948.
[66] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[67] H. Liu, M. Wang, Low-frequency Fourier spectrum for predicting membrane protein types, Biochem. Biophys. Res. Commun. 336 (2005) 737–739.
[68] W. Chen, P.M. Feng, H. Lin, ISS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed. Res. Int. 2014 (2014) 623149.
[69] Y. Xu, X. Wen, X.J. Shao, IHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594–7610.
[70] S. Wan, M.W. Mak, S.Y. Kung, GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition, J. Theor. Biol. 323 (2013) 40–48.
[71] G.S. Han, Z.G. Yu, V. Anh, A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC, J. Theor. Biol. 344 (2014) 31–39.
[72] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[73] N. Cristianini, J. Shawe-Taylor, An Introduction of Support Vector Machines and Other Kernel–based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.
[74] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[75] N. Cristianini, J. Shawe-Taylor, Kernel-induced feature spaces, in: An Introduction to Support Vector Machines and Other Kernel–based Learning Methods, Cambridge University Press, Cambridge, UK, 2000, chap. 3.
[76] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (2007) 423–428.

[77] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.

[78] K.C. Chou, Z.C. Wu, X. Xiao, ILoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2012) 629–641.

[79] W.Z. Lin, J.A. Fang, X. Xiao, ILoc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. BioSyst. 9 (2013) 634–644.

[80] L. Chen, W.M. Zeng, Y.D. Cai, Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities, PLoS ONE 7 (2012) e35254.

[81] X. Xiao, P. Wang, W.Z. Lin, J.H. Jia, IAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.

[82] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[83] K.C. Chou, H.B. Shen, Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, Nat. Sci. 2 (2013) 1090–1103.

[84] Y.K. Chen, K.B. Li, Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 318 (2013) 1–12.

[85] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition, J. Theor. Biol. 360 (2014) 109–116.

[86] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.

[87] H.B. Shen, J. Yang, Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids 33 (2007) 57–67.

[88] K.C. Chou, S. Forsen, Graphical rules for enzyme-catalyzed rate laws, Biochem. J. 187 (1980) 829–835.

[89] Z.C. Wu, X. Xiao, 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, J. Theor. Biol. 267 (2010) 29–34.

[90] I.W. Althaus, J.J. Chou, A.J. Gonzales, F.J. Kezdy, W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E, Biochemistry 32 (1993) 6548–6554.

[91] K.C. Chou, Graphic rule for drug metabolism systems, Curr. Drug Metab. 11 (2010) 369–378.

[92] G.P. Zhou, The disposition of the LZCC protein residues in Wenxiang diagram provides new insights into the protein–protein interaction mechanism, J. Theor. Biol. 284 (2011) 142–148.

[93] J.A. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2005) 861–874.