Analytical Biochemistry 490 (2015) 26–33

Contents lists available at ScienceDirect

Analytical Biochemistry

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

iRNA-Methyl: Identifying N^6-methyladenosine sites using pseudo nucleotide composition

Wei Chen [a, b, *], Pengmian Feng [c], Hui Ding [c], Hao Lin [b, d, **], Kuo-Chen Chou [b, e, ***]

[a] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China
[b] Gordon Life Science Institute, Belmont, MA 02478, USA
[c] School of Public Health, North China University of Science and Technology, Tangshan 063000, China
[d] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 13 July 2015
Received in revised form
13 August 2015
Accepted 16 August 2015
Available online 24 August 2015

Keywords:
RNA methylation
Pseudo dinucleotide composition: PseKNC
Global sequence pattern
Flexible scaled window

ABSTRACT

Occurring at adenine (A) with the consensus motif GAC, N^6-methyladenosine (m^6A) is one of the most abundant modifications in RNA, which plays very important roles in many biological processes. The nonuniform distribution of m^6A sites across the genome implies that, for better understanding the regulatory mechanism of m^6A, it is indispensable to characterize its sites in a genome-wide scope. Although a series of experimental technologies have been developed in this regard, they are both time-consuming and expensive. With the avalanche of RNA sequences generated in the postgenomic age, it is highly desired to develop computational methods to timely identify their m^6A sites. In view of this, a predictor called "iRNA-Methyl" is proposed by formulating RNA sequences with the "pseudo dinucleotide composition" into which three RNA physiochemical properties were incorporated. Rigorous cross-validation tests have indicated that iRNA-Methyl holds very high potential to become a useful tool for genome analysis. For the convenience of most experimental scientists, a web-server for iRNA-Methyl has been established at http://lin.uestc.edu.cn/server/iRNA-Methyl by which users can easily get their desired results without needing to go through the mathematical details.

More than 100 kinds of post-transcriptional RNA modifications have been found in eukaryotic messenger RNA (mRNA) [1]. Among these modifications, N^6-methyladenosine (m^6A) is the most abundant one that is also the first RNA reversible one [2]. As shown in Fig.1, the modification occurs on the sixth nitrogen atom of adenine. Because it was found during the 1970s, m^6A has been identified in all three kingdoms of life [3–6] and is associated with a number of biological processes, including mRNA splicing, export, stability, and immune tolerance [7–9].

With the development of high-throughput techniques such as MeRIP-Seq [10] and m^6A-seq [11], the genome-wide distribution of m^6A is now available for several species such as *Saccharomyces cerevisiae* [12], *Mus musculus* [13], and *Homo sapiens* [13]. These experimental results revealed that m^6A sites tend to occur near the stop codon, in 3′ UTR, and within long internal exons [10,13]. The nonrandom distribution of m^6A sites across the genome is highly conserved from yeast to human, suggesting that m^6A modification is both fundamental and important for organisms [12,13]. The current biochemical methods are, however, both costly and time-consuming in performing genome-wide analysis. Therefore, it is in high demand to develop computational methods for analyzing the distribution and function of m^6A so as to help speed up the genome-wide m^6A detection.

Abbreviations: mRNA, messenger RNA; m^6A, N^6-methyladenosine; NAC, nucleic acid composition; PseDNC, pseudo dinucleotide composition; SVM, support vector machine; RBF, radial basis kernel function; PseAAC, pseudo amino acid composition; PseKNC, pseudo k-tupler nucleotide composition.

* Corresponding author. Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China.
** Corresponding author. University of Electronic Science and Technology of China, Chengdu 610054, China.
*** Corresponding author. Gordon Life Science Institute, Belmont, MA 02478, USA.
E-mail addresses: chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.ab.2015.08.021
0003-2697/

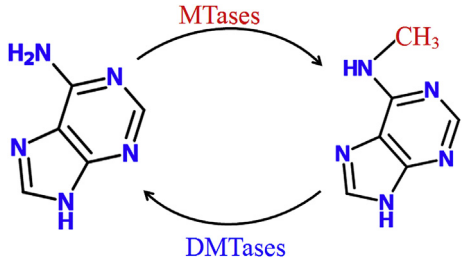

Fig.1. Illustration showing the N^6-methylation and demethylation of adenosine. The formation of m^6A is catalyzed by N^6-adenosyl methyltransferases (MTases), whereas its reversible modification (demethylation) is catalyzed by demethyltransferases (DMTases).

Unfortunately, to our best knowledge, so far there is no computational tool available whatsoever for detecting m^6A. In view of this, the current study was initiated in an attempt to develop a new computational predictor by which one can easily identify m^6A sites.

As demonstrated by a series of recent publications [14–22], to establish a really useful sequence-based statistical predictor for a biological system and also to make the presentation logically crystal clear, we should follow the five-step guidelines [23]: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy; and (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we elaborate how to deal with these steps one by one.

Materials and methods

Benchmark dataset

Because the sites of m^6A in the *S. cerevisiae* genome share a consensus motif GAC where its center base has the potential to be methylated [12], for facilitating description later, we use the following scheme to represent an RNA sample:

$$\mathbf{R}_\xi(\mathbb{GAC}) = N_{-\xi}N_{-(\xi-1)}\cdots N_{-2}N_{-1}\mathbb{GAC}N_{+1}N_{+2}\cdots N_{+(\xi-1)}N_{+\xi}, \tag{1}$$

where the center $\mathbb{A}$ represents "adenine," the subscript ξ is an integer, $N_{-\xi}$ represents the ξ-th upstream nucleotide from the center, $N_{+\xi}$ represents the ξ-th downstream nucleotide, and so forth (Fig.2). The $(2\xi + 3)$-tuple RNA sample $\mathbf{R}_\xi(\mathbb{GAC})$ can be further classified into the following categories:

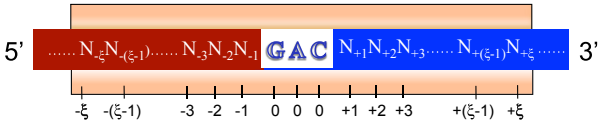

Fig.2. Schematic drawing showing how to use the flexible scaled window along an RNA sequence to collect the potential m^6A-containing segments. See Eqs. (1)–(5) and the relevant text for further explanation.

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

where $\mathbf{R}_\xi^+(\mathbb{GAC})$ denotes a true methylation segment with adenine at its center, $\mathbf{R}_\xi^-(\mathbb{GAC})$ denotes a false methylation segment with adenine at its center, and the symbol $\in$ means "a member of" in the set theory.

As elaborated in 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 because the outcome obtained in this way is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset $\mathbb{S}_\xi$ for the current study can be formulated as

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

where the positive subset $\mathbb{S}_\xi^+$ contains only the samples of true methylation segments $\mathbf{R}_\xi^+(\mathbb{GAC})$ and the negative subset $\mathbb{S}_\xi^-$ contains only the samples of false methylation segments $\mathbf{R}_\xi^-(\mathbb{GAC})$ (see Eq. (2)), whereas $\bigcup$ represents the symbol for "union" in the set theory.

Because the length of RNA sample $\mathbf{R}_\xi(\mathbb{GAC})$ is $2\xi + 3$ (see Eq. (1)), the benchmark dataset with different ξ value will contain RNA segments with different number of nucleotides, as illustrated below:

$$\text{The length of RNA samples in } \mathbb{S}_\xi = \begin{cases} 21 \text{ nucleotides}, & \text{if } \xi = 9 \\ 31 \text{ nucleotides}, & \text{if } \xi = 14 \\ 41 \text{ nucleotides}, & \text{if } \xi = 19 \\ 51 \text{ nucleotides}, & \text{if } \xi = 24 \\ 61 \text{ nucleotides}, & \text{if } \xi = 29 \\ \vdots & \vdots \end{cases}. \tag{4}$$

Preliminary tests had indicated, however, that best prediction results were achieved when $\xi = 24$. Accordingly, hereafter we focus on the RNA samples with 51 nucleotides only.

The detailed procedures to construct $\mathbb{S}_{\xi=24}$ are as follows. First, as done in Ref. [25], slide the $(2\xi + 3) = 51$-tuple nucleotide window along each of the RNA sequences taken from *S. cerevisiae* genome, and collected were only those RNA segments that have GAC at the center and A (adenine) or G (guanine) at the position of N_{-1} (see Eq. (1)); this is done because the consensus motif for m^6A determined by experiments for *S. cerevisiae* genome is RGAC (R = A/G) [12]. Second, if the upstream or downstream in an RNA was less than $\xi = 24$ or greater than $L - 24$ (L is the RNA's length), the lacking nucleotide was filled with its mirror image (Fig.3). Third, the RNA segment samples obtained in this way were put into the positive

(A) Mirror image for 5′ terminus

$$N_{-1}N_{-2}\cdots N_{-22}N_{-23} \Leftrightarrow N_{-23}N_{-22}\cdots N_{-2}N_{-1}$$

(B) Mirror image for 3′ terminus

$$N_{L-23}N_{L-22}\cdots N_{L-1}N_L \Leftrightarrow N_LN_{L-1}\cdots N_{L-22}N_{L-23}$$

Fig.3. Schematic illustration showing the mirror image of the 5′ RNA terminal segment (A) and the 3′ RNA terminal segment (B). The symbol $\Leftrightarrow$ represents a mirror. The real RNA segment is colored in blue, whereas its mirror image is colored in red. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

subset $\mathbb{S}_{\xi}^{+}$ if their centers have been experimentally annotated as the methylation sites; otherwise, they were put into the negative subset $\mathbb{S}_{\xi}^{-}$. Fourth, using the 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-conflicted (i.e., simultaneously occurring in both methylation subset $\mathbb{S}_{\xi}^{+}$ and nonmethylation subset $\mathbb{S}_{\xi}^{-}$).

By following the aforementioned five steps, we first obtained a benchmark dataset consisting of 1307 positive samples and 33,280 negative samples. It is a very imbalanced dataset in which the size of $\mathbb{S}_{\xi}^{-}$ is overwhelmingly greater than that of $\mathbb{S}_{\xi}^{+}$. To minimize the underprediction or overprediction [27] caused by such a highly skewed benchmark dataset, we randomly picked out 1307 ones from the 33,280 negative samples to form the negative dataset $\mathbb{S}_{\xi}^{-}$.

The detailed sequences for the 1307 positive samples and 1307 negative samples are given in the online Supplementary material. They can also be downloaded at http://lin.uestc.edu.cn/server/iRNAMethyl/data.

Representation of RNA samples

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

$$\mathbf{R} = N_1 N_2 N_3 \cdots N_i \cdots N_{51}, \tag{5}$$

where N_1 represents the first nucleotide at the sample sequence position 1, N_2 represents the second nucleotide at the position 2, and so forth. They can be any of the four nucleotides, that is,

$$N_i \in \{\, \mathrm{A(adenine)} \;\; \mathrm{C\,(cytosine)} \;\; \mathrm{G\,(guanine)} \;\; \mathrm{U\,(uracil)} \,\}. \tag{6}$$

Based on the sequential model of Eq. (5), one could directly use BLAST [28] to perform statistical analysis. Unfortunately, this kind of straightforward and intuitive approach failed to work when a query RNA sequence sample did not have significant similarity to any of the character-known RNA sequences.

To deal with this problem, investigators could not help but resort to the discrete or vector model. Actually, an important reason for them to do so is that all of the existing machine-learning algorithms can be directly used to handle vector models but not sequences, as elaborated in Ref. [29].

The most simple vector model for an RNA sequence is its nucleic acid composition (NAC), that is,

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

where f(A), f(C), f(G), and f(U) are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G), and uracil (U) in the RNA sequence, respectively; the symbol **T** is the transpose operator. As we can see from Eq. (7), however, if using NAC to represent a RNA sample, all of its sequence order information would be completely lost.

If using the k-tuple nucleotide (k-mer) composition to represent the RNA sequence, the corresponding vector will have a dimension of 4^k. With the incensement of k values, the vector's dimension will increase rapidly, leading to the so-called "high-dimension disaster" [30] or overfitting problem that will significantly reduce the deviation tolerance or cluster-tolerant capacity [31] so as to lower the prediction success rate or stability. Therefore, the k-mer approach is useful only when the value of k is very small. In other words, it can be used only to incorporate the local or short-range sequence order or pattern information, but certainly not the global or long-range sequence order or pattern information.

To approximately cover the long-range sequence pattern information, one popular and well-known method is to use the pseudo component approach originally proposed for dealing with protein/peptide sequences [32]. Ever since being introduced in 2001, the approach and its concept have been penetrating to nearly all of the areas of computational proteomics (see, e.g., Refs. [33–40] as well as a long list of articles cited in a recent review article [41]. Because the pseudo component approach has been widely and increasingly used, some publicly accessible web-servers [42–44] have been established, allowing users to generate various kinds of pseudo components according to their needs to study many different problems in computational proteomics. Recently, the concept of pseudo component approach was further extended to study the problems in computational genetics and genomics [18,21,45,46]. Meanwhile, the corresponding web-servers have been developed accordingly for generating various kinds of pseudo components for DNA sequences [47–49] and RNA sequences [50,51].

To incorporate both the local and global sequence pattern information of the RNA sequences, we adopted the approach of pseudo 2-tuple nucleotide composition or pseudo dinucleotide composition (PseDNC), that is, representing the RNA sample of Eq. (5) with the formulation below:

$$\mathbf{R} = [d_1\; d_2\; \cdots\; d_{16}\; d_{16+1}\; \cdots\; d_{16+\lambda}]^{\mathbf{T}}, \tag{8}$$

where,

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 16) \\ \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < u \le 16+\lambda) \end{cases}. \tag{9}$$

In Eq. (9), f_u (u = 1,2,⋯,16) is the normalized occurrence frequency of the u-th non-overlapping dinucleotides in the RNA sequence, and

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} \mathrm{C}_{i,\,i+j}\; (j = 1, 2, \cdots, \lambda;\; \lambda < L), \tag{10}$$

where θ_1 is called the first-tier correlation factor that reflects the sequence order correlation between all of the most contiguous dinucleotide along a RNA sequence (Fig.4A), θ_2 is the second-tier correlation factor between all of the second-most contiguous dinucleotide (Fig.4B), θ_3 is the third-tier correlation factor between all of the third-most contiguous dinucleotide (Fig.4C), and so forth.

Now, it is clear that the first 16 components in Eq. (8) are used to incorporate the short-range or local sequence order information of the RNA sample, whereas the remaining components are used for its long-range or global sequence order information. Obviously, λ can also be viewed as the number of the total pseudo components used to reflect the long-range or global sequence effect [50,51] and w of Eq. (9) is the weight factor [32,35]. The concrete values for $\boldsymbol{\lambda}$ and w are further discussed later.

In Eq. (10), the coupling factor $\mathrm{C}_{i,i+j}$ is given by

$$\mathrm{C}_{i,\,i+j} = \frac{1}{\mu}\sum_{g=1}^{\mu}\left[P_g(\mathrm{D}_i) - P_g(\mathrm{D}_{i+j})\right]^2, \tag{11}$$

where μ is the number of RNA physicochemical properties considered that is equal to 3 in the current study and is further explained below.

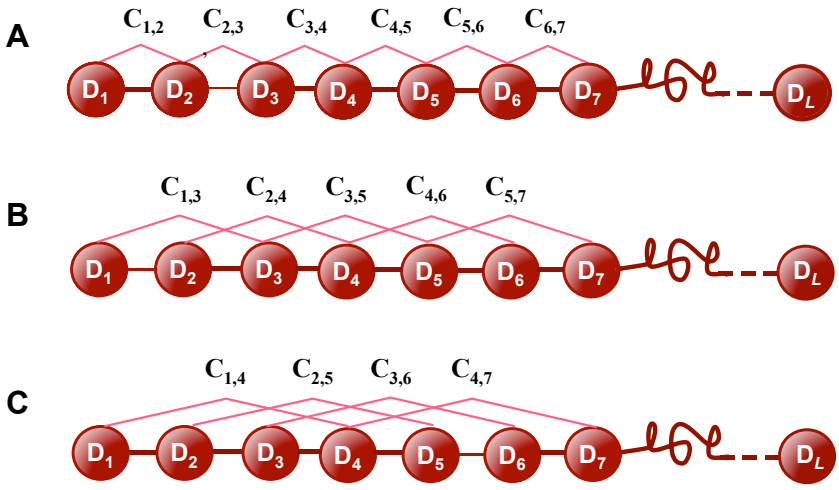

Fig.4. Schematic illustration showing the correlations of dinucleotides along an RNA sequence. (A) The first-tier correlation reflects the sequence order mode between all of the most contiguous non-overlapping dinucleotide. (B) The second-tier correlation reflects the sequence order mode between all of the second-most contiguous non-overlapping dinucleotide. (C) The third-tier correlation reflects the sequence order mode between all of the third-most contiguous non-overlapping dinucleotide.

Table 1
Original values of the three physicochemical properties for the 16 different dinucleotides in RNA.

Dinucleotide	Enthalpy (K_a/mol)	Entropy (eU)	Free energy (K_a/mol)
GG	−12.2	−29.7	−3.26
GA	−13.3	−35.5	−2.35
GC	−14.2	−34.9	−3.42
GU	−10.2	−26.2	−2.24
AG	−7.6	−19.2	−2.08
AA	−6.6	−18.4	−0.93
AC	−10.2	−26.2	−2.24
AU	−5.7	−15.5	−1.10
CG	−8.0	−19.4	−2.36
CA	−10.5	−27.8	−2.11
CC	−12.2	−29.7	−3.26
CU	−7.6	−19.2	−2.08
UG	−7.6	−19.2	−2.11
UA	−8.1	−22.6	−1.33
UC	−10.2	−26.2	−2.35
UU	−6.6	−18.4	−0.93

Note. See text for further explanation.

RNA property parameters

Because the formation of RNA secondary structure will decrease the m^6A methylation [52], three physicochemical properties—enthalpy [53], entropy [53], and free energy [54]—that can quantify the RNA secondary structures [55–57] are used to calculate the global or long-range sequence order effects via Eqs. (10) and (11). The concrete values of these three physicochemical properties are given in Table 1. Note that before substituting them into Eq. (11), all of the original values $P_g(\mathrm{D_i})$ (i = 1,2,3) were subjected to a standard conversion, as described by the following equation:

$$P_g(\mathrm{D}_i) \Leftarrow \frac{P_g(\mathrm{D}_i) - \langle P_g(\mathrm{D}_i)\rangle}{\mathrm{SD}\{\langle P_g(\mathrm{D}_i)\rangle\}}, \tag{12}$$

where the symbol ⟨⟩ means taking the average of the quantity therein over the 16 different dinucleotides, and SD means the corresponding standard deviation. For the detailed mathematical formulation of SD, see Eq. (4) of the original article [32] or Eq. (4) of the 2005 article [35]. The advantage to do so is that the converted values obtained by Eq. (12) will have a zero mean value over the 16 different dinucleotides and will remain unchanged if going through the same conversion procedure again [24].

Support vector machine

Support vector machine (SVM) is a machine-learning algorithm based on the statistical learning theory. It has been widely used in the realm of bioinformatics (see, e.g., [16]; [19]; [45]; [46]). Its basic principle is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space by using the following decision function:

$$F(\vec{X}) = \mathrm{sgn}\left\{\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{X}, \vec{X}_i) + b\right\}, \tag{13}$$

where α_i is the Lagrange multipliers, b is the offset, $\vec{X}$ is the query input vector, $\vec{X}_i$ is the i-th training vector, y_i represents the type of the i-th training vector, $K(\vec{X}, \vec{X}_i)$ is a kernel function that defines an inner product in a high-dimensional feature space, and sgn is the sign function. Due to its effectiveness and speed in the nonlinear classification process, the radial basis kernel function (RBF) was used in the current study. For a brief formulation of SVM and how it works, see the article Ref. [58]; for more details about SVM, see the monograph Ref. [59].

The package LIBSVM 2.84 (http://www.csie.ntu.edu.tw/~cjlin) written by Chang and Lin was employed to perform SVM in the current study. The SVM algorithm contains two parameters; one is the regularization parameter C, and the other is the kernel width parameter γ. In the current study, the two parameters were determined by an optimization procedure in which the grid search and 10-fold cross-validation were performed. The final results obtained in this way were $C = 32$ and $\gamma = 0.0078125$.

The predictor obtained via the above procedures is called "iRNA-Methyl."

Results and discussion

Metrics used to evaluate the prediction quality

The current study is a kind of binary classification problem, that is, for a given RNA sample, whether it is a positive one (belonging to the methylation segment) or a negative one (belonging to the nonmethylation segment). For this kind of binary classification problem, the following set of metrics was often used to measure the prediction quality:

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{MCC} = \dfrac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases}, \tag{14}$$

where TP is the true positive, TN is the true negative, FP is the false positive, FN is the false negative, Sn is the sensitivity, Sp is the specificity, Acc is the accuracy, and MCC is the Matthews correlation coefficient [60]. The metrics formulated in Eq. (14) is not easy to understand for most experimental scientists, and hence here we prefer to use the following formulation as done by many investigators in a series of recent publications (see, e.g., Refs. [14,17,22,61–65]):

$$\begin{cases} \mathrm{Sn} = 1-\dfrac{N_-^+}{N^+} & 0\le \mathrm{Sn}\le 1 \\ \mathrm{Sp} = 1-\dfrac{N_+^-}{N^-} & 0\le \mathrm{Sp}\le 1 \\ \mathrm{Acc} = \Lambda = 1-\dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0\le \mathrm{Acc}\le 1 \\ \mathrm{MCC} = \dfrac{1-\left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1+\dfrac{N_+^- - N_-^+}{N^+}\right)\left(1+\dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1\le \mathrm{MCC}\le 1 \end{cases}, \tag{15}$$

where N^+ is the total number of positive samples or true methylation RNA segments investigated, N_-^+ is the number of true methylation RNA samples incorrectly predicted to be false methylation segments, N^- is the total number of negative samples or nonmethylation RNA samples investigated, and N_+^- is the number of nonmethylation RNA samples incorrectly predicted to be methylation segments. According to Eq. (15), the following is crystal clear. When $N_-^+ = 0$, meaning that none of the positive sample was incorrectly predicted to be negative, we have sensitivity Sn = 1. When $N_-^+ = N^+$, meaning that all of the positive samples were incorrectly predicted to be negative, we have sensitivity Sn = 0. Likewise, when $N_+^- = 0$, meaning that none of the negative samples was mispredicted, we have specificity Sp = 1. When $N_+^- = N^-$, meaning that all of the negative samples were incorrectly predicted as positive, we have specificity Sp = 0. When $N_-^+ = N_+^- = 0$, meaning that none of the samples in the positive dataset and none of the samples in the negative dataset were incorrectly predicted, we have overall accuracy Acc = 1 and MCC = 1. When $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all of the samples in the positive dataset and all of the samples in the negative dataset were incorrectly predicted, we have overall accuracy Acc = 0 and MCC = −1. When $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have Acc = 0.5 and MCC = 0, meaning no better than random guessing. As we can see from the above discussion, it would make the meanings of sensitivity, specificity, overall accuracy, and Matthews correlation coefficient much more intuitive and easier to understand by using the formulation of Eq. (15), particularly for the meaning of MCC.

It should be pointed out, however, that the set of metrics as defined in Eq. (14) or Eq. (15) is valid only for the single-label systems. For the multi-label systems, whose emergence has become more frequent in system biology [66–69] and system medicine [29,70], a completely different set of metrics as defined in Ref. [27] is needed.

Method used to conduct cross-validation

With a set of clearly defined metrics available to measure the prediction quality, the next thing is what validation method should be used to derive the metrics values. In statistical prediction, the following three cross-validation methods are often used to derive the metrics values for a predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [71]. Of the three methods, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset, as elucidated in Ref. [23] and demonstrated by Eqs. 28–32 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. [37,39,40,72–79]).

Accordingly, in this study we also used the jackknife test to evaluate the accuracy of the current predictor. During the jackknife test, each of the samples in the benchmark dataset is in turn singled out as an independent test sample and all of the rule parameters are calculated without including the sample being identified. Although the jackknife test may take more computational time, it is worthwhile because it will always yield a unique outcome for a given benchmark dataset.

Parameter determination and anticipated success rates

As we can see from Eqs. (9) and (10), the current model depends on the two parameters w and λ. The former is the weight factor usually within the range from 0 to 1, whereas the latter is the number of correlation tiers considered to reflect the global sequence pattern effect (Fig.4). Generally speaking, the greater the λ is, the more global sequence pattern information the model contains. But if λ is too large, it would reduce the cluster-tolerant capacity [31] so as to lower the cross-validation accuracy due to overfitting or the "high-dimension disaster" problem [30]. Therefore, our searching for the optimal values of the two parameters was within the ranges given below:

$$\begin{cases} 3\le \lambda\le 6 & \text{with step } \Delta = 1 \\ 0.1\le w\le 1 & \text{with step } \Delta = 0.1 \end{cases}. \tag{16}$$

At this step, for reducing computational time, the iRNA-Methyl predictor was examined by the 10-fold cross-validation on the benchmark dataset $\mathbb{S}$ (see Eq. (3) as well as the Supplementary material). The results obtained in this way are illustrated in Fig.5, from which we can see that, when $\lambda = 6$ and $w = 0.9$, the predictor's accuracy (Acc) reaches its peak, indicating that the optimal λ and w values for the proposed predictor are 6 and 0.9, respectively, when trained by the current benchmark dataset.

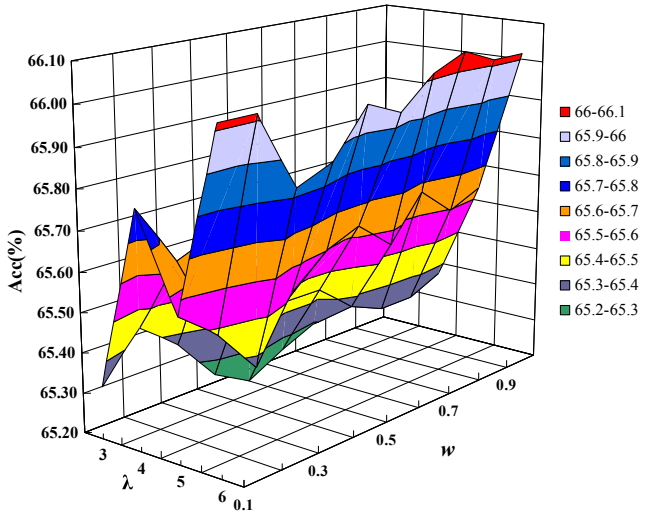

Fig.5. Three-dimensional graph showing the accuracies obtained in the 10-fold cross-validation with different values of w and λ.

Subsequently, with λ fixed at 6 and w fixed at 0.9, the rigorous jackknife tests were performed to calculate the Sn, Sp, Acc, and MCC as defined in Eq. (15) for the iRNA-Methyl predictor on the same benchmark dataset. The results obtained in this way are listed in Table 2. Before the availability of iRNA-Methyl, to predict the methylation sites in an RNA sample, one could not help but use the sequence-similarity-search-based tools (e.g., BLAST [28]) to search for those character-known sequences with high similarity to the query sample. According to the BLAST approach, the query sample will be predicted as the true methylation RNA segment if it is most similar to the samples in the positive subset; otherwise, it will be predicted as the false methylation RNA segment. Although it was quite straightforward and intuitive, unfortunately, the BLAST approach failed to work when the query sample did not have significant similarity to any of the character-known sequences as elucidated in Ref. [24]. With the availability of iRNA-Methyl, however, one can easily get the desired results via its web-server. The success rates obtained by iRNA-Methyl and the BLAST approach via the rigorous jackknife tests on the same benchmark dataset are given in Table 2, from which we can see the following. First, for the rates obtained by the BLAST approach, there is a big gap between Sn and Sp, indicating that the predicted results by the BLAST approach are very unstable with quite low specificity; in contrast, the corresponding rates obtained by iRNA-Methyl are much more even. Second, the Acc rate achieved by iRNA-Methyl is approximately 10% higher than that of the BLAST approach, and the MCC rate of iRNA-Methyl is two times that of BLAST, indicating that the iRNA-Methyl predictor is superior to the BLAST approach not only in overall accuracy but also in stability.

Table 2
Comparison of iRNA-Methyl with the other method in identifying methylation sites in RNA.

Prediction method	Sn (%)	Sp (%)	Acc (%)	MCC
iRNA-Methyl[a]	70.55	60.63	65.59	0.29
BLAST approach[b]	71.76	38.79	55.27	0.11

Note. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient.
[a] Proposed in this article.
[b] Based on the sequence similarity principle [28].

All of this implies that the iRNA-Methyl predictor proposed in this article is quite promising and may become a useful high-throughput tool in identifying m^6A sites.

Web-server and guide for users

For the convenience of most experimental scientists, a publicly accessible web-server for iRNA-Methyl has been established. Moreover, to maximize users' convenience, a step-by-step guide on how to use it to get the desired results is given below:

Step 1 Open the web-server at http://lin.uestc.edu.cn/server/iRNA-Methyl and you will see the top page of the iRNA-Methyl predictor on your computer screen, as shown in Fig.6. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.
Step 2 Either type or copy/paste the query RNA sequences into the input box at the center of Fig.6. The input sequence should be in FASTA format. For examples of RNA sequences in FASTA format, click 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 RNA sequences in the Example window as the input, you will see the following on the screen of your computer. (1) RNA sequence 1 contains 5 GAC (with adenine at its middle) consensus motifs, of which only those at the sequence position 128 is predicted to be the methylation sites or m^6A site, whereas all of the others are not. (2) RNA sequence 2 contains 8 GAC consensus motifs, of which only those at the sequence position 332 is predicted to be the methylation sites, whereas all of the others are not. All of these results are fully consistent with the experimental observations.
Step 4 Click on the Data button to download the datasets used to train and test the model.
Step 5 Click on the Citation button to find the relevant article that documents the detailed development and algorithm of iRNA-Methyl.

Conclusions

Encouraged by the successes of pseudo amino acid composition (PseAAC) in dealing with protein/peptide sequences, a new

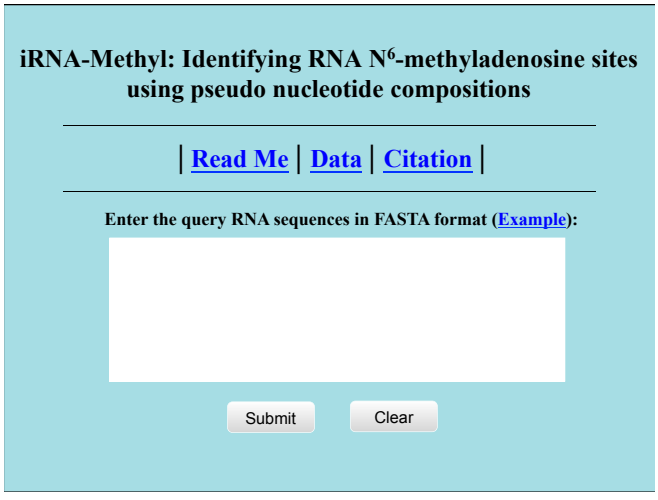

Fig.6. Semi-screenshot showing the top page of the iRNA-Methyl web-server. Its website address is http://lin.uestc.edu.cn/server/iRNA-Methyl.

predictor called iRNA-Methyl has been proposed for identifying m^6A sites in the *S. cerevisiae* genome by incorporating the global and long-range sequence pattern information of RNA via the pseudo k-tupler nucleotide composition (PseKNC) approach. The jackknife test on a rigorous benchmark dataset demonstrates that the iRNA-Methyl predictor is very promising.

Although the current iRNA-Methyl was trained by the benchmark dataset derived from *S. cerevisiae* genome, it can be extended to analyze the genomes of other species as well if trained by the benchmark datasets from those species.

In particular, it has not escaped our notice that the current approach and its mathematical frame can also be used to develop different computational predictors for identifying various other modification sites in RNA.

A user-friendly web-server for iRNA-Methyl has been established at http://lin.uestc.edu.cn/server/iRNA-Methyl by which users can easily obtain their desired results without the need to go through the complicated mathematics involved, which was presented here just for its integrity. It is anticipated that iRNA-Methyl may become a useful high-throughput tool for conducting genome analysis.

Acknowledgments

The authors thank the three anonymous reviewers, whose constructive comments were very helpful in strengthening the presentation of this study. This work was supported by the National Natural Science Foundation of China (61100092, 61202256, and 61301260), the Natural Science Foundation of Hebei Province (C2013209105), the Foundation of Science and Technology Department of Hebei Province (132777133), the Applied Basic Research Program of Sichuan Province (2015JY0100), the Fundamental Research Funds for the Central Universities, China (ZYGX2013J102), and the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (BJ2014028).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ab.2015.08.021.

References

[1] W.A. Cantara, P.F. Crain, J. Rozenski, J.A. McCloskey, K.A. Harris, X. Zhang, F.A. Vendeix, D. Fabris, P.F. Agris, The RNA modification database, RNAMDB: 2011 update, Nucleic Acids Res. 39 (2011) D195–D201.
[2] J. Liu, G. Jia, Methylation modifications in eukaryotic messenger RNA, J. Genet. Genomics 41 (2014) 21–33.
[3] C.M. Wei, A. Gershowitz, B. Moss, 5′-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA, Biochemistry 15 (1976) 397–401.
[4] R. Levis, S. Penman, 5′-Terminal structures of poly(A)$^+$ cytoplasmic messenger RNA and of poly(A)$^+$ and poly(A)$^-$ heterogeneous nuclear RNA of cells of the dipteran *Drosophila melanogaster*, J. Mol. Biol. 120 (1978) 487–515.
[5] J.L. Nichols, "Cap" structures in maize poly(A)-containing RNA, Biochim. Biophys. Acta 563 (1979) 490–495.
[6] M.J. Clancy, M.E. Shambaugh, C.S. Timpte, J.A. Bokar, Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N^6-methyladenosine in mRNA: a potential mechanism for the activity of the *IME4* gene, Nucleic Acids Res. 30 (2002) 4509–4518.
[7] K. Kariko, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA, Immunity 23 (2005) 165–175.
[8] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.G. Yang, C. He, N^6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO, Nat. Chem. Biol. 7 (2011) 885–887.
[9] T.W. Nilsen, Molecular biology: internal mRNA methylation finally finds functions, Science 343 (2014) 1207–1208.
[10] K.D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C.E. Mason, S.R. Jaffrey, Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons, Cell 149 (2012) 1635–1646.
[11] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of N^6-methyladenosine by m^6A-seq based on immunocapturing and massively parallel sequencing, Nat. Protoc. 8 (2013) 176–189.
[12] S. Schwartz, S.D. Agarwala, M.R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T.S. Mikkelsen, R. Satija, G. Ruvkun, S.A. Carr, E.S. Lander, G.R. Fink, A. Regev, High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis, Cell 155 (2013) 1409–1421.
[13] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, G. Rechavi, Topology of the human and mouse m^6A RNA methylomes revealed by m^6A-seq, Nature 485 (2012) 201–206.
[14] W. Chen, H. Lin, P.M. Feng, C. Ding, Y.C. Zuo, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (2012) e47843.
[15] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, 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] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (2015) e0121501.
[22] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.
[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] H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (2007) 633–640.
[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] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.
[28] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.
[29] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[30] T. Wang, J. Yang, H.B. Shen, Predicting membrane protein types by the LLDA algorithm, Protein Pept. Lett. 15 (2008) 915–921.
[31] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216–224.
[32] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (Erratum: 2001, vol. 44, p. 60).
[33] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.
[34] 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.
[35] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[36] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE–ACM Trans. Comput. Biol. Bioinform. 9 (2012) 467–475.
[37] 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.

[38] M. Hayat, N. Iqbal, Discriminating protein structure classes by incorporating pseudo average chemical shift to Chou's general PseAAC and support vector machine, Comput. Methods Programs Biomed. 116 (2014) 184–192.
[39] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.
[40] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.
[41] W. Chen, H. Lin, K.C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed. Res. Int. 2014 (2014) 623149.
[47] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[48] W. Chen, X. Zhang, J. Brooker, H. Lin, PseKNC-general: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (2015) 119–120.
[49] B. Liu, F. Liu, L. Fang, X. Wang, repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, Bioinformatics 31 (2015) 1307–1309.
[50] B. Liu, F. Liu, L. Fang, repRNA: a web server for generating various feature vectors of RNA sequences, Mol. Genet. Genomics (2015), http://dx.doi.org/10.1007/s00438-015-1078-7.
[51] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.
[52] P. Narayan, R.L. Ludwiczak, E.C. Goodwin, F.M. Rottman, Context effects on N^6-adenosine methylation sites in prolactin mRNA, Nucleic Acids Res. 22 (1994) 419–426.
[53] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, D.H. Turner, Improved free-energy parameters for predictions of RNA duplex stability, Proc. Natl. Acad. Sci. U. S. A. 83 (1986) 9373–9377.
[54] T. Xia, J. SantaLucia Jr., M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, D.H. Turner, Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs, Biochemistry 37 (1998) 14719–14735.
[55] J.L. Fiore, B. Kraemer, F. Koberling, R. Edmann, D.J. Nesbitt, Enthalpy-driven RNA folding: single-molecule thermodynamics of tetraloop–receptor tertiary interaction, Biochemistry 48 (2009) 2550–2558.
[56] T.I. Shaw, A. Manzour, Y. Wang, R.L. Malmberg, L. Cai, Analyzing modular RNA structure reveals low global structural entropy in microRNA sequence, J. Bioinform. Comput. Biol. 9 (2011) 283–298.
[57] D.H. Mathews, D.H. Turner, Prediction of RNA secondary structure by free energy minimization, Curr. Opin. Struct. Biol. 16 (2006) 270–278.
[58] Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[59] N. Cristianini, J. Shawe-Taylor, An Introduction of Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.
[60] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (2007) 423–428.
[61] 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.
[62] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.
[63] 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.
[64] 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. 33 (2015) 1731–1742.
[65] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, 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.
[66] 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.
[67] 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.
[68] X. Xiao, Z.C. Wu, iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, J. Theor. Biol. 284 (2011) 42–51.
[69] X. Wang, W. Zhang, Q. Zhang, G.Z. Li, MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier, Bioinformatics 31 (2015) 2639–2645.
[70] X. Xiao, P. Wang, W.Z. Lin, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.
[71] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[72] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins 44 (2001) 57–59.
[73] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.
[74] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.
[75] B. Liu, L. Fang, F. Liu, X. Wang, K.C. Chou, iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach, J. Biomol. Struct. Dyn. (2015), http://dx.doi.org/10.1080/07391102.2015.1014422.
[76] J.L. Min, X. Xiao, K.C. Chou, iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking, Biomed. Res. Int. 2013 (2013) 701317.
[77] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, K.C. Chou, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.
[78] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, K.C. Chou, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015) 2221–2233.
[79] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, K.C. Chou, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.08.025.