Analytical Biochemistry 497 (2016) 60–67

Contents lists available at ScienceDirect

Analytical Biochemistry

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

pRNAm-PC: Predicting N^6-methyladenosine sites in RNA sequences via physical–chemical properties

Zi Liu [a, b], Xuan Xiao [a, c, d, *], Dong-Jun Yu [b], Jianhua Jia [a], Wang-Ren Qiu [a], Kuo-Chen Chou [d, e]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[c] Information School, ZheJiang Textile and Fashion College, NingBo 315211, China
[d] Gordon Life Science Institute, Boston, MA 02478, USA
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form
2 December 2015
Accepted 23 December 2015
Available online 31 December 2015

Keywords:
N^6-Methyldenosine sites
Auto-covariance
Cross covariance
pRNAm-PC
Pseudo dinucleotide composition

ABSTRACT

Just like PTM or PTLM (post-translational modification) in proteins, PTCM (post-transcriptional modification) in RNA plays very important roles in biological processes. Occurring at adenine (A) with the genetic code motif (GAC), N^6-methyldenosine (m^6A) is one of the most common and abundant PTCMs in RNA found in viruses and most eukaryotes. Given an uncharacterized RNA sequence containing many GAC motifs, which of them can be methylated, and which cannot? It is important for both basic research and drug development to address this problem. Particularly with the avalanche of RNA sequences generated in the postgenomic age, it is highly demanded to develop computational methods for timely identifying the N^6-methyldenosine sites in RNA. Here we propose a new predictor called pRNAm-PC, in which RNA sequence samples are expressed by a novel mode of pseudo dinucleotide composition (PseDNC) whose components were derived from a physical–chemical matrix via a series of auto-covariance and cross covariance transformations. It was observed via a rigorous jackknife test that, in comparison with the existing predictor for the same purpose, pRNAm-PC achieved remarkably higher success rates in both overall accuracy and stability, indicating that the new predictor will become a useful high-throughput tool for identifying methylation sites in RNA, and that the novel approach can also be used to study many other RNA-related problems and conduct genome analysis. A user-friendly Web server for pRNAm-PC has been established at http://www.jci-bioinfo.cn/pRNAm-PC, by which users can easily get their desired results without needing to go through the mathematical details.

Introduction

Post-transcriptional modifications (PTCM) of RNA play a crucial role in understanding various RNA metabolisms, such as messenger RNA (mRNA) stability, splicing export, immune tolerance, and transcription [1–4]. So far, more than 100 distinct PTCMs have been identified in mRNA, transfer RNA (tRNA) and ribosomal RNA (rRNA) [5,6]. Among these modification, N^6-methyldenosine (m^6A) is the one of the most important PTCMs, which is catalyzed by a methyltransferase complex containing at least one subunit of METTL3 (methyltransferase like 3), and the process is reversible under the catalysis of demethylases FTO and ALKBH5, as shown in Fig. 1.

RNA N^6-methyldenosine has various biological functions. It is important for cell fate determination in yeast [7,8], and it is also significant for embryo development in plants [9]. Accordingly, knowledge of mRNA m^6A sites is vitally important for both basic research and drug development.

Narayan et al. [10] investigated the distribution of m^6A in mRNA by means of the experimental techniques such as TLC (thin layer chromatography) and HPLC (high performance liquid chromatography). Recently, it was observed using various experimental high-throughput experimental tools such as CHIP-Seq [11,12] and MeRIP-Seq [13,14] that the m6A sites are not randomly distributed, but are near the stop codons and within the coding sequences [11]. Recently, Harcour et al. [15] used the selective polymerase approach to detect N(6)-methyladenosine in RNA.

* Corresponding author. Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China.
E-mail address: xxiao@gordonlifescience.org (X. Xiao).

http://dx.doi.org/10.1016/j.ab.2015.12.017
0003-2697/

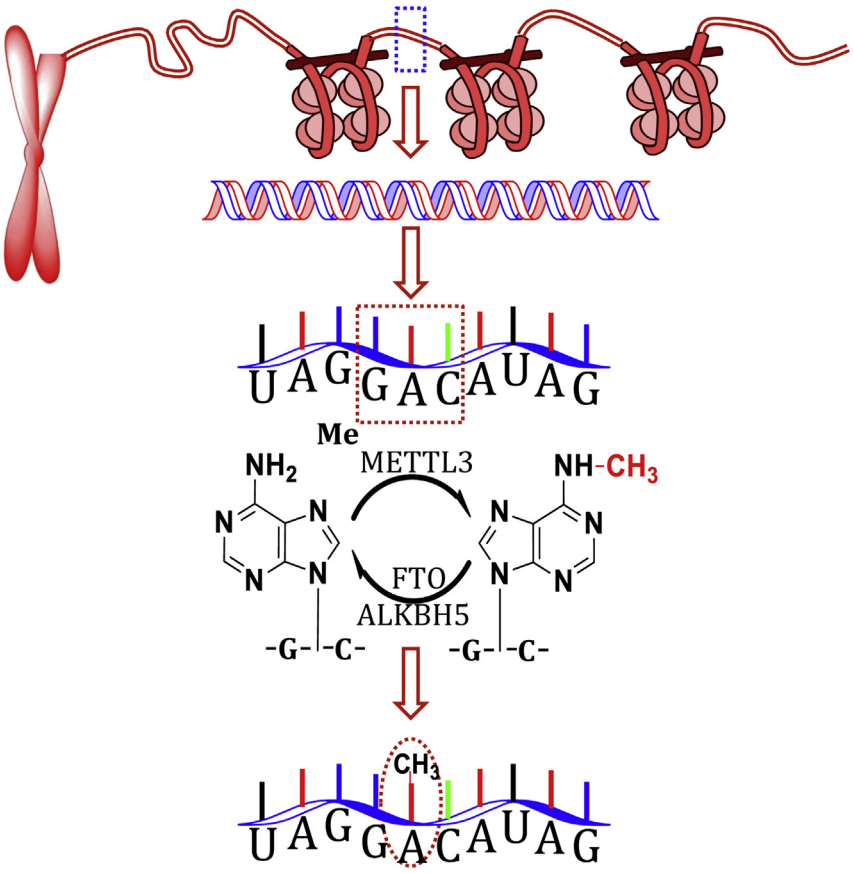

Fig.1. An illustration of reversible N^6-methylation and demethylation in mRNA. The formation of m^6A is catalyzed by methyltransferase METTL3, and its reversible modification (demethylation) is catalyzed by demethylases FTO and ALKBH5.

The studies by these authors are significant in stimulating the development of this area. But it is both time-consuming and expensive to determine m^6A sites in RNA using purely experimental techniques. Particularly, with the avalanche of RNA sequences occurring in the postgenomic age, it is highly demanded to develop computational tools for rapidly determining the methylation sites in RNA. Actually, in a pioneer work, Chen et al. [16] developed a computational method to predict m^6A sites in RNA via pseudo nucleotide composition [17,18] or PseKNC [19,20], a strategy for extending pseudo amino acid composition or PseAAC [21] in dealing with protein/peptide sequences to treat DNA/RNA sequences. For users' convenience, their method also has a Web-server called "iRNA-Methyl," which is the first computational tool ever established for predicting the m^6A sites in RNA. According to their report, however, its overall success rate is 65.59%, implying that further efforts are needed to enhance its accuracy.

The present study was devoted to this problem. According to Chou's five-step rule [22] and fulfilled in a series of recent publications [23–30], to establish a really useful sequence-based statistical predictor for a biological system, we need to consider the following five procedures: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the biological sequence samples with an effective mathematical expression that can effectively correlate with the target to be predicted; (3) introduce or develop a powerful algorithm (or operation engine) to calculate the prediction; (4) properly carry out cross validation tests to objectively evaluate the anticipated accuracy; (5) establish a user-friendly Web server accessible to the public. Below, we describe how to fulfill these steps one by one.

Materials and methods

Benchmark dataset

In literature the benchmark dataset usually consists of a training dataset and a testing dataset: the former is used for training a model, while the latter used for testing the model. But as pointed out in a comprehensive review [31], there is no need to artificially separate a benchmark dataset into the two parts if the prediction model is examined by the jackknife test or subsampling (*K*-fold) cross validation, because the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset set $\mathbb{S}$ for the current study can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-, \tag{1}$$

where the positive subset $\mathbb{S}^+$ contains only the samples of methylation RNA segments; the negative subset $\mathbb{S}^-$ contains only the samples of nonmethylation RNA segment; and $\cup$ represents the symbol for "union" in set theory. The detailed samples were downloaded from the iRNA-Methyl Web server [16]. Such a high-quality benchmark dataset for studying N^6-methyldenosine sites in RNA can also be obtained from Ref. [32]. As shown in Supporting Information S1, the benchmark dataset consists of 1307 positive samples and 1307 negative samples, each being of a 51-tuple nucleotide. The reason each of such samples contains a 51-tuple nucleotide is that they were collected by sliding a flexible window [33] along

each of the RNA sequences taken from the *Saccharomyces cerevisiae* genome. The window's width is $(2\xi + 3)$; when $\xi = 24$ or $(2\xi\ 3) = 51$, best prediction results were achieved, as elaborated in Ref. [32].

From RNA sequential model to discrete model

An RNA sample in $\mathbb{S}$ of Eq. (1) or Supporting Information S1 can generally be expressed as

$$\mathbf{R} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_{51}, \tag{2}$$

where N_1 represents the first nucleotide of the RNA sample at its sequence position 1, N_2 the second nucleotide at its position 2, and so forth. They can be any of the four nucleotides; i.e.,

$$\mathrm{N}_i \in \{\ \mathrm{A}\ (\text{adenine})\quad \mathrm{C}\ (\text{cytosine})\quad \mathrm{G}\ (\text{guanine})\quad \mathrm{U}\ (\text{uracil})\ \}, \tag{3}$$

where $i = 1, 2, \cdots, 51$ and the symbol $\in$ means "a member of" in set theory.

Based on the sequential model of Eq. (2), one could directly utilize BLAST [34] to perform statistical analysis. Unfortunately, this kind of straightforward and intuitive approach failed to work when a query RNA sample did not have significant similarity to any of the character-known RNA sequences.

To overcome this problem, investigators have shifted their focus to the discrete or vector model. The advantage of doing so is also due to the fact that nearly all the existing machine-learning algorithms can be used directly to handle vector models but not sequences, as elaborated in Ref. [35].

The simplest vector model for an RNA sequence is its nucleic acid composition (NAC); i.e.,

$$\mathbf{R} = [f(\mathrm{A})\ f(\mathrm{C})\ f(\mathrm{G})\ f(\mathrm{U})]^{\mathbf{T}}, \tag{4}$$

where $f(\mathrm{A})$, $f(\mathrm{C})$, $f(\mathrm{G})$, and $f(\mathrm{U})$ are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G), and uracil (U) in the RNA sequence, respectively; the symbol $\mathbf{T}$ is the transpose operator. As we can see from Eq. (4), however, if NAC were used to represent a RNA sample, all its sequence order information would be completely lost.

If the RNA sequence sample is represented by the k-tuple nucleotide (or k-mers) composition [17], the corresponding feature vector will contain 4^k components, as given by

$$\mathbf{R} = [\,f_1 \quad f_2 \quad f_3 \quad \cdots \quad f_i \quad \cdots \quad f_{4^k}\,]^{\mathbf{T}}, \tag{5}$$

where f_i represents the normalized occurrence frequency of the ith k-mer. As we can see from Eq. (5), when $k > 4$ the number of the vector components will rapidly increase, causing the so-called "high-dimension disaster" [36] or overfitting problem, which will significantly reduce the deviation tolerance or cluster-tolerant capacity [37], lowering the prediction success rate or stability. Therefore, the k-mers approach is useful only when the value of k is very small. In other words, it can only be used to incorporate the local or short-range sequence-order information, but certainly not the global or long-range sequence-order information. To approximately cover the long-range sequence-order effects, one popular and well-known method is to use the pseudo components that were originally introduced in dealing with protein/peptide sequences [21,38] and recently extended to deal with DNA/RNA sequences [17–20,23,39–42].

According to the concept of pseudo components Ref. [22], the RNA sequence can be generally formulated.

$$\mathbf{R} = [\Psi_1\ \Psi_2\ \cdots\ \Psi_u\ \cdots\ \Psi_\Omega]^{\mathbf{T}}, \tag{6}$$

where the subscript Ω is an integer and its value, as well as the components Ψ_u $(u = 1, 2, \cdots, \Omega)$, will depend on how the desired information is extracted from the RNA sequence of Eq. (2).

Below, we use the "physical–chemical property matrix" and "auto-covariance and cross covariance transformations" to define the Ω elements in Eq. (6).

Physical–chemical property matrix

There are $4 \times 4 = 16$ different dinucleotides or dimers in an RNA sequence, i.e., AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU (cf. Eq. (3)). Each of the sixteen dimers has a different set of physical–chemical (PC) properties. Thus, an RNA sample can be encoded by a series of PC values. In the current study, the following 10 PC properties were considered: (1) PC^1: rise [43]; (2) PC^2: roll [43]; (3) PC^3: shift [43]; (4) PC^4: slide [43]; (5) PC^5: tilt [43]; (6) PC^6: twist [43]; (7) PC^7: enthalpy [44]; (8) PC^8: entropy [45]; (9) PC^9: stack energy [43]; (10) PC^{10}: free energy [45]. Listed in Table 1 are their original values, based on which the RNA sample in Eq. (2) can be converted to a $10 \times (51-1) = 10 \times 50$ physical–chemical property matrix,

$$\mathbb{PC} = \begin{bmatrix} \mathrm{PC}^1(\mathrm{N}_1\mathrm{N}_2) & \mathrm{PC}^1(\mathrm{N}_2\mathrm{N}_3) & \cdots & \mathrm{PC}^1(\mathrm{N}_{50}\mathrm{N}_{51}) \\ \mathrm{PC}^2(\mathrm{N}_1\mathrm{N}_2) & \mathrm{PC}^2(\mathrm{N}_2\mathrm{N}_3) & \cdots & \mathrm{PC}^2(\mathrm{N}_{50}\mathrm{N}_{51}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{PC}^{10}(\mathrm{N}_1\mathrm{N}_2) & \mathrm{PC}^{10}(\mathrm{N}_2\mathrm{N}_3) & \cdots & \mathrm{PC}^{10}(\mathrm{N}_{50}\mathrm{N}_{51}) \end{bmatrix}, \tag{7}$$

where $\mathrm{PC}^j(\mathrm{N}_i\mathrm{N}_{i+1})$ is the jth $(j = 1, 2, \cdots, 10)$ PC value for the $\mathrm{N}_i\mathrm{N}_{i+1}$ dinucleotide in Eq. (2).

Before the data of Table 1 were substituted into Eq. (7), however, they were subject to a standard conversion through the following equation [31],

$$y_m = [x_m - \mathrm{mean}(x)]/\mathrm{std}(x), \tag{8}$$

where x_m stands for the original PC value in Table 1 for the mth $(m = 1, 2, \cdots, 10)$ dinucleotide, $\mathrm{mean}(x)$ for the average score for the 16 dinucleotides, and $\mathrm{std}(x)$ for the corresponding standard deviation. Listed in Table 2 are the corresponding converted values of y_m, which will have a zero mean value over the 16 dinucleotides or dimers, and will remain unchanged if they go through the same conversion procedure again.

Auto-covariance and cross covariance

In statistics, the auto-covariance is the covariance of a stochastic process against a parameter-shift version of itself, while the cross covariance is used to refer to the covariance between two random vectors. In this study, we use the two concepts of covariance to transform the matrix of Eq. (7) to a length-fixed feature vector.

According to the concept of auto-covariance (AC), the correlation of the same PC property between two subsequences separated by λ dinucleotides or dimers can be expressed as

$$\mathrm{AC}(m,\lambda) = \frac{\sum_{j=1}^{50-\lambda}\left[\mathrm{PC}^m(\mathrm{N}_j\mathrm{N}_{j+1}) - \overline{\mathrm{PC}^m}\right]\left[\mathrm{PC}^m(\mathrm{N}_{j+\lambda}\mathrm{N}_{j+1+\lambda}) - \overline{\mathrm{PC}^m}\right]}{50-\lambda} \quad (m = 1, 2, \cdots, 10), \tag{9}$$

Table 1
Original values of the 10 physico-chemical properties for each of the 16 dinucleotides.

Code	Dimer	PC^1	PC^2	PC^3	PC^4	PC^5	PC^6	PC^7	PC^8	PC^9	PC^{10}
1	AA	3.18	7.0	−0.08	−1.27	−0.8	31	−6.82	−18.4	−13.7	−0.9
2	AC	3.24	4.8	0.23	−1.43	0.8	32	−11.40	−26.2	−13.8	−2.1
3	AG	3.3	8.5	−0.04	−1.50	0.5	30	−10.48	−19.2	−14	−1.7
4	AU	3.24	7.1	−0.06	−1.36	1.1	33	−9.38	−15.5	−15.4	−0.9
5	CA	3.09	9.9	0.11	−1.46	1.0	31	−10.44	−27.8	−14.4	−1.8
6	CC	3.32	8.7	−0.01	−1.78	0.3	32	−13.39	−29.7	−11.1	−2.9
7	CG	3.3	12.1	0.3	−1.89	−0.1	27	−10.64	−19.4	−15.6	−2
8	CU	3.3	8.5	−0.04	−1.50	0.5	30	−10.48	−19.2	−14.0	−1.7
9	GA	3.38	9.4	0.07	−1.70	1.3	32	−12.44	−35.5	−14.2	−2.3
10	GC	3.22	6.1	0.07	−1.39	0.0	35	−14.88	−34.9	−16.9	−3.4
11	GG	3.32	12.1	−0.01	−1.78	0.3	32	−13.39	−29.7	−11.1	−2.9
12	GU	3.24	4.8	0.23	−1.43	0.8	32	−11.40	−26.2	−13.8	−2.1
13	UA	3.26	10.7	−0.02	−1.45	−0.2	32	−7.69	−22.6	−16.0	−1.1
14	UC	3.38	9.4	0.07	−1.70	1.3	32	−12.44	−26.2	−14.2	−2.1
15	UG	3.09	9.9	0.11	−1.46	1.0	31	−10.44	−19.2	−14.4	−1.7
16	UU	3.18	7.0	−0.08	−1.27	−0.8	31	−6.82	−18.4	−13.7	−0.9

Table 2
The converted data obtained from Table 1 via Eq. (8).

Code	Dimer	PC^1	PC^2	PC^3	PC^4	PC^5	PC^6	PC^7	PC^8	PC^9	PC^{10}
1	AA	−0.83	−0.67	−1.13	1.34	−1.84	−0.26	1.72	0.95	0.29	1.35
2	AC	−0.14	−1.64	1.50	0.49	0.54	0.34	−0.27	−0.31	0.23	−0.26
3	AG	0.55	0.00	−0.79	0.12	0.09	−0.86	0.13	0.82	0.10	0.28
4	AU	−0.14	−0.62	−0.96	0.87	0.98	0.93	0.61	1.42	−0.83	1.35
5	CA	−1.87	0.62	0.48	0.33	0.83	−0.26	0.15	−0.57	−0.17	0.14
6	CC	0.78	0.09	−0.53	−1.36	−0.20	0.34	−1.13	−0.88	2.02	−1.34
7	CG	0.55	1.60	2.09	−1.95	−0.80	−2.65	0.06	0.79	−0.97	−0.13
8	CU	0.55	0.00	−0.79	0.12	0.09	−0.86	0.13	0.82	0.10	0.28
9	GA	1.47	0.40	0.14	−0.94	1.28	0.34	−0.72	−1.82	−0.04	−0.53
10	GC	−0.37	−1.07	0.14	0.71	−0.65	2.13	−1.77	−1.72	−1.83	−2.01
11	GG	0.78	1.60	−0.53	−1.36	−0.20	0.34	−1.13	−0.88	2.02	−1.34
12	GU	−0.14	−1.64	1.50	0.49	0.54	0.34	−0.27	−0.31	0.23	−0.26
13	UA	0.09	0.98	−0.62	0.39	−0.95	0.34	1.34	0.27	−1.23	1.08
14	UC	1.47	0.40	0.14	−0.94	1.28	0.34	−0.72	−0.31	−0.04	−0.26
15	UG	−1.87	0.62	0.48	0.33	0.83	−0.26	0.15	0.82	−0.17	0.28
16	UU	−0.83	−0.67	−1.13	1.34	−1.84	−0.26	1.72	0.95	0.29	1.35

where λ is an integer within the range from 0 to 49, and $\overline{PC^m}$ is the mean of the data along the mth row in the matrix of Eq. (7), as given by

$$\overline{PC^m} = \frac{\sum_{j=1}^{50} PC^m(N_jN_{j+1})}{50}. \tag{10}$$

As we can see from Eq. (9), by means of the auto-covariance approach, we can generate $10 \times \lambda$ components associated with the physical–chemical properties of an RNA sample in Eq. (2).

On the other hand, according to the concept of cross covariance (CC), the correlation between two subsequences each belonging to a different PC property can be formulated by

$$CC(\mu_1, \mu_2, \lambda) = \frac{\sum_{j=1}^{50-\lambda}\left[PC^{\mu_1}(N_jN_{j+1}) - \overline{PC^{\mu_1}}\right]\left[PC^{\mu_2}(N_{j+\lambda}N_{j+1+\lambda}) - \overline{PC^{\mu_2}}\right]}{50-\lambda} \quad (\mu_1 = 1, 2, \cdots, 10;\ \mu_2 = 1, 2, \cdots, 10;\ \mu_1 \neq \mu_2), \tag{11}$$

indicating that the cross-covariance approach can generate $10 \times 9 \times \lambda$ components associated with the sample of Eq. (2).

With Eqs. (9) and (11), a total of $(10 \times \lambda + 10 \times 9 \times \lambda) = 100 \times \lambda$ components were generated by auto-covariance and cross covariance via 10 different physical–chemical properties. Preliminary tests indicated, however, that the outcomes were most promising when $\lambda = 4$. Therefore, the RNA sample is hereafter formulated by

$$\mathbf{R} = [\Psi_1\ \Psi_2\ \cdots\ \Psi_u\ \cdots\ \Psi_{400}]^{\mathbf{T}}, \tag{12}$$

where Ψ_u is the uth of the 400 components generated by Eqs. (9)–(10) as described above.

Support vector machine

SVM is a machine-learning algorithm based on statistical learning theory. It has been widely used in the realm of bioinformatics (see, e.g., [16,23–28,30,41,42,46–53]). The basic idea of SVM is to construct a separating hyperplane to maximize the margin between the positive dataset and negative dataset. The nearest two points to the hyperplane are called the support vectors. SVM first constructs a hyperplane based on the training dataset, and then maps an input vector from the input space into a vector in a higher-

dimensional Hilbert space, where the mapping is determined by a kernel function. A trained SVM can output a class label (in our case, methylation or nonmethylation) based on the mapping vector of the input vector.

For a brief formulation of SVM and how it works, see the papers [54,55]; for more details about SVM, see the monograph [56].

In the current study, the RNA samples as formulated by Eq. (12) were used as inputs for the SVM classifier. Given a query RNA sample, the classifier can quite accurately predict which class it belongs to after training by a relevant dataset, i.e., clearly indicating whether it is a "methylation RNA segment" or "nonmethylation RNA segment." Also, the LIBSVM algorithm [57] was employed, which is software for SVM classification and regression. The kernel function was set as the radial basis function (RBF) and its two parameters were optimized with the benchmark dataset $\mathbb{S}$ via a two-dimensional grid search (Fig. 2) performed by LIBSVM [57]. The optimized parameters thus obtained in the current study were $C = 16$ and $\gamma = 0.0039$.

The predictor obtained via the aforementioned procedures is called pRNAm-PC, where "p" stands for "predict," "RNAm" for "RNA methylation site," and "PC" for "physical–chemical properties."

Results and discussion

As mentioned in the Introduction, one of the important procedures in developing a new predictor is properly and objectively evaluating its quality [22], which actually consists of the following two aspects: what metrics should be used to quantitatively measure the prediction accuracy, and what kind of test method should be utilized to derive the metrics' values. Below, we address these problems.

A set of more intuitive metrics to define the prediction quality

To facilitate quantitative analysis, in this study we used a set of more intuitive and easier-to-understand metrics formulated in terms of the symbols introduced by Chou [58] in studying signal peptide prediction. According to Chou's formulation, the sensitivity Sn, specificity Sp, overall accuracy Acc, and Matthews correlation coefficient MCC can be expressed [59,60].

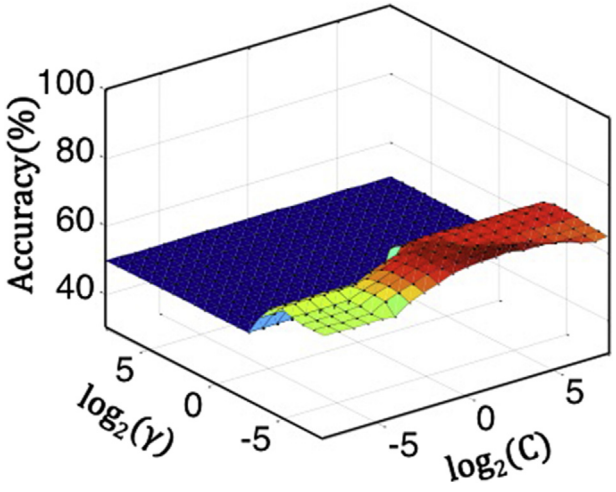

Fig.2. Three-dimensional plot to show how to find the optimal values of C and γ via a two-dimensional grid search.

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1\,, \end{cases} \tag{13}$$

where N^+ is the total number of the positive samples or true methylation RNA segments investigated while N_-^+ is the number of true methylation RNA samples incorrectly predicted to be of false methylation segment; N^- the total number of the negative samples or nonmethylation RNA samples investigated while N_+^- is the number of the nonmethylation RNA samples incorrectly predicted to be of methylation segment.

According to Eq. (13), the following is crystal clear. When $N_-^+ = 0$, meaning none of the positive sample was incorrectly predicted to be a negative one, we have the sensitivity Sn = 1. When $N_-^+ = N^+$, meaning that all the positive samples were incorrectly predicted to be the negative, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$, meaning none of the negative samples was mispredicted, we have the specificity Sp = 1; whereas $N_+^- = N^-$, meaning that all the negative samples were incorrectly predicted as positive, we have the 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 the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all the samples in the positive dataset and all the samples in the negative dataset were incorrectly predicted, we have the overall accuracy Acc = 0 and MCC = −1; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have Acc = 0.5 and MCC = 0, meaning no better than a random guess.

As we can see from the above discussion, using the metrics formulated in Eq. (13) rather than the conventional formulation would make the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and clearer, particularly for the meaning of MCC, as concurred by a series of recent publications [16,23–25,27–29,41,42,49–53,61–65].

But note that the set of metrics in Eq. (13) is valid only for the single-label systems. For the multilabel systems, whose emergence has become more frequent in system biology [66–69] and system medicine [70], a completely different set of metrics is needed as elaborated in Ref. [71].

Cross validation

In statistical prediction, the following three cross validation methods are often used to calculate the values of the four metrics in Eq. (13) for a predictor: independent dataset test, subsampling (or K-fold cross validation) test, and jackknife test [72]. Of the three methods, the jackknife test is deemed the least arbitrary, which can always yield a unique outcome for a given benchmark dataset as elucidated in Ref. [22] and demonstrated by Eqs. (28)–(32) therein. Therefore, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., [73–77]).

Accordingly, in this study we also use 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 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.

Comparison with the existing predictor

The success rates achieved by the pRNAm-PC predictor via the jackknife test on the benchmark dataset of Eq. (1) (cf. the Supporting Information S1) are given in Table 3. To facilitate comparison, listed there are also the corresponding rates achieved by iRNA-Methyl [16], the only peer counterpart in the area of predicting the methylation sites in RNA. As we can see from Table 3, the new predictor pRNAm-PC proposed in this paper remarkably outperformed its counterpart, particularly in Acc and MCC; the former stands for the overall accuracy, and the latter for the stability.

Graphs are a useful vehicle for studying complicated biological systems because they can provide intuitive insights, as demonstrated by a series of previous studies (see, e.g., [78–84]). To provide an intuitive comparison, the graph of receiver operating characteristic (ROC) [85,86] was adopted to show the improvement of pRNAm-PC over iRNA-Methyl. The blue graphic line (in the web version) in Fig. 3 is the ROC curve for the iRNA-Methyl predictor, while the red graphic line is that for the proposed predictor pRNAm-PC. The area under the ROC curve is called the AUC (area under the curve). The greater the AUC value is, the better the predictor will be [85,86]. As we can see from Fig. 3, the area under the red curve is remarkably greater than that under the blue one, indicating that the proposed predictor is indeed better than iRNA-Methyl [16], the only existing bioinformatics tool for identifying the methylation sites in RNA. Therefore, we anticipate that pRNAm-PC may become a useful tool in this important area, or at the very least, play a complementary role to the existing method.

Web server and its user guide

To enhance the value of its practical applications, a Web server for pRNAm-PC has been established. Furthermore, to maximize the convenience of most experimental scientists, a step-by-step guide is provided, by which users can easily get their desired results without the need to go through the detailed mathematical equations involved in this paper.

Step 1. Opening the Web server at http://www.jci-bioinfo.cn/pRNAm-PC, you will see the top page of pRNAm-PC on your computer screen, as shown in Fig. 4. Click on the Read Me button to see a brief introduction to the predictor.

Table 3
A comparison[a] of the pRNAm-PC predictor with the other existing method for predicting methylation sites in RNA.

Predictor	ACC (%)	MCC (%)	Sn (%)	Sp (%)	ROC (%)
iRNA-Methyl[b]	65.59	0.29	70.55	60.63	70.48
pRNAm-PC[c]	69.74	0.40	69.72	69.75	76.28

[a] The results listed below were obtained by the jackknife test on a same benchmark dataset (cf. Eq. (1)).
[b] See Ref. [16].
[c] Proposed in this paper.

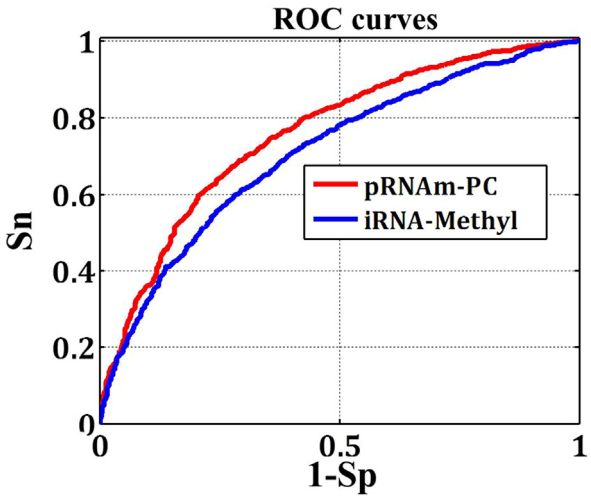

Fig.3. The ROC curves to show the predictor's quality.

Step 2. Either type or copy/paste the query RNA sequences into the input box at the center of Fig. 4. The input sequence should be in the FASTA format. For examples of 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, in about 20 s after submitting, you will see the following on the screen of your computer: (1) Sequence-1 contains 205 nucleic acid residues, of which only the one (highlighted with red) at sequence position 128 is predicted to be the methylation site; all the others are not. (2) Sequence-2 contains 271 residues, of which 7 are predicted as methylation sites; they are located at sequence positions 6, 12, 26, 62, 77, 246, and 157, as highlighted with red. All these results are fully consistent with the experimental observations.
Step 4. As shown in the lower panel of Fig. 4, you may also choose batch prediction by entering your e-mail address and your desired batch input file (in FASTA format of course) via the Browse button. To see the sample of a batch input file, click the button Batch-example. After clicking the button Batch-submit, 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 in this study for training and testing the predictor.
Step 6. Click the Citation button to find the relevant papers that document the detailed development and algorithm of pRNAm-PC.

Conclusions

The distribution of N^6-methyladenosine (m^6A) sites in RNA is important for in-depth understanding of its regulatory mechanism, and for drug development as well. Among the existing high-throughput tools for characterizing the m^6A sites in a genome-wide scope, pRNAm-PC is the most powerful one. It has not escaped our notice that the approaches introduced here, such as using the pseudo dinucleotide composition to represent RNA samples and defining the pseudo components via a physical–chemical matrix of 2-tuple nucleotides, can also be used to address many other important problems in genome analysis.

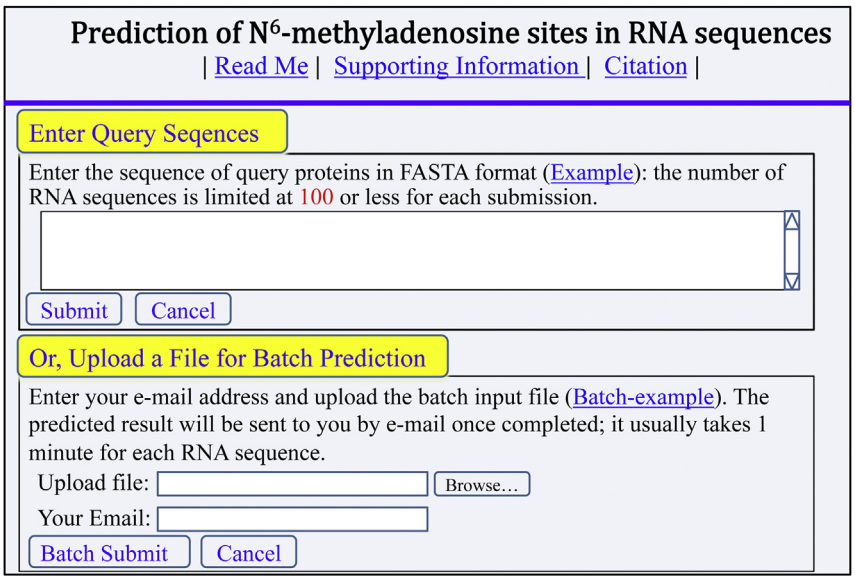

Fig.4. A semi-screenshot of the top page of the Web server pRNAm-PC at http://www.jci-bioinfo.cn/pRNAm-PC.

Acknowledgments

The authors wish to thank the two anonymous reviewers for their constructive comments, which were very helpful in strengthening the presentation of this paper. This work was partially supported by the National Nature Science Foundation of China (Nos. 31260273, 61261027, 31560316, 61373062), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 20120BDH80023), the Natural Science Foundation of Jiangxi Province, China (Nos. 20114BAB211013, 20122BAB211033, 20122BAB201044, 20122BAB201020) the Department of Education of JiangXi Province (GJJ12490, GJJ14640), the LuoDi plan of the Department of Education of JiangXi Province (KJLD12083), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ab.2015.12.017.

References

[1] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.G. Yang, C. He, N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO, Nat. Chem. Biol. 7 (2011) 885–887.
[2] K. Karikó, 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.
[3] T.W. Nilsen, Internal mRNA methylation finally finds functions, Science 343 (2014) 1207–1208.
[4] Y. Niu, Z. Xu, Y.S. Wu, M.M. Li, X.J. Wang, Y.G. Yang, N6-methyl-adenosine (m6A) in RNA: an old modification with a novel epigenetic function, Genomics Proteomics Bioinforma 11 (2013) 8–17.
[5] 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.
[6] D. Globisch, D. Pearson, A. Hienzsch, T. Brückl, M. Wagner, I. Thoma, P. Thumbs, V. Reiter, A.C. Kneuttinger, M. Müller, Systems-based analysis of modified tRNA bases, Angew. Chem. Int. Ed. 50 (2011) 9739–9742.
[7] M.J. Clancy, M.E. Shambaugh, C.S. Timpte, J.A. Bokar, Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene, Nucleic Acids Res. 30 (2002) 4509–4518.
[8] S.D. Agarwala, H.G. Blitzblau, A. Hochwagen, G.R. Fink, RNA methylation by the MIS complex regulates a cell fate decision in yeast, 2012.
[9] S. Zhong, H. Li, Z. Bodi, J. Button, L. Vespa, M. Herzog, R.G. Fray, MTA is an arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor, Plant Cell 20 (2008) 1278–1288.
[10] P. Narayan, R.L. Ludwiczak, E.C. Goodwin, F.M. Rottman, Context effects on N6-adenosine methylation sites in prolactin mRNA, Nucleic Acids Res. 22 (1994) 419–426.
[11] D. Dan, M.M. Sharon, S. Schraga, S.D. Mali, U. Lior, O. Sivan, C. Karen, J.H. Jasmine, A. Ninette, K. Martin, Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq, Nature 485 (2012) 201–206.
[12] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of N6-methyladenosine by m 6A-seq based on immunocapturing and massively parallel sequencing, Nat. Protoc. 8 (2013) 176–189.
[13] Y. Saletore, K. Meyer, J. Korlach, I.D. Vilfan, S. Jaffrey, C.E. Mason, The birth of the epitranscriptome: deciphering the function of RNA modifications, Genome Biol. 13 (2012) 1078–1084.
[14] J. Meng, Z. Lu, H. Liu, L. Zhang, S. Zhang, Y. Chen, M.K. Rao, Y. Huang, A protocol for RNA methylation differential analysis with MeRIP-Seq data and exome peak R/bioconductor package, Methods 69 (2014) 274–281.
[15] E.M. Harcourt, E. Thomas, P.J. Batista, H.Y. Chang, E.T. Kool, Identification of a selective polymerase enables detection of N(6)-methyladenosine in RNA, J. Am. Chem. Soc. 135 (2013) 19079–19082.
[16] W. Chen, P. Feng, H. Ding, iRNA-methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33 (also, Data in Brief, 2015, 5: 376-378).
[17] W. Chen, T.Y. Lei, D.C. Jin, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[18] 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.
[19] 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.
[20] B. Liu, F. Liu, X. Wang, J. Chen, 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.
[21] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255. Structure, Function, and Genetics (Erratum: ibid., 2001, Vol. 44, 60).
[22] 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.
[23] 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.
[24] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels, BioMed Res. Int. (BMRI) 2014 (2014) 286419.
[25] H. Lin, E.Z. Deng, H. Ding, 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.
[26] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. (IJMS) 15 (2014) 1746–1766.

[27] B. Liu, L. Fang, S. Wang, X. Wang, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.
[28] R. Xu, J. Zhou, B. Liu, Y.A. He, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. (JBSD) 33 (2015) 1720–1730.
[29] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC), J. Biomol. Struct. Dyn. (2015). http://dx.doi.org/10.1080/07391102.2015.1095116.
[30] B. Liu, L. Fang, R. Long, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics (2015). http://dx.doi.org/10.1093/bioinformatics/btv604.
[31] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[32] W. Chen, P. Feng, H. Ding, Benchmark data for identifying N^6-methyl-adenosine sites in the *Saccharomyces cerevisiae* genome, Data Brief 5 (2015) 376–378.
[33] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. (BBRC) 357 (2007) 633–640.
[34] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.
[35] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[36] T. Wang, J. Yang, H.B. Shen, Predicting membrane protein types by the LLDA algorithm, Protein Peptide Lett. 15 (2008) 915–921.
[37] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. (BBRC) 264 (1999) 216–224.
[38] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[39] W. Chen, X. Zhang, J. Brooker, PseKNC-general: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (2015) 119–120.
[40] 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.
[41] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed. Res. Int. (BMRI) 2014 (2014) 623149.
[42] Z. Liu, X. Xiao, W.R. Qiu, iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77 (also, Data in Brief, 2015, 4: 87-89).
[43] P. Alberto, N. Agnes, L. Filip, L.F. Javier, O. Modesto, The relative flexibility of B-DNA and A-RNA duplexes: database analysis, Nucleic Acids Res. 32 (2004) 6144–6151.
[44] J.R. Goñi, A. Pérez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol. 8 (2007) R263.
[45] 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, in: Proceedings of the National Academy of Sciences of the United States of America 83, 1986, pp. 9373–9377.
[46] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.
[47] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.
[48] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[49] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. (IJMS) 15 (2014) 4915–4937.
[50] 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. (IJMS) 15 (2014) 7594–7610.
[51] 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.
[52] 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. (JBSD) 33 (2015) 1731–1742.
[53] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. (JBSD) 33 (2015) 2221–2233.
[54] 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.
[55] 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.
[56] N. Cristianini, J. Shawe-Taylor, An Introduction of Support Vector Machines and Other Kernel-based Learning Methodds, Cambridge University Press, Cambridge, UK, 2000.
[57] C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[58] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[59] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[60] 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.
[61] 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. (BMRI) 2014 (2014) 947416.
[62] 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.
[63] 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.
[64] B. Liu, L. Fang, F. Liu, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (2015) e0121501.
[65] 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. 34 (2016) 223–235. http://dx.doi.org/10.1080/07391102.2015.1014422.
[66] 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.
[67] W.Z. Lin, J.A. Fang, 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, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.
[72] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[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] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins Struct. Funct. Genet. 44 (2001) 57–59.
[75] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.
[76] H.B. Shen, Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells, Biopolymers 85 (2007) 233–240.
[77] M. Mandal, A. Mukhopadhyay, U. Maulik, Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC, Med. Biol. Eng. Comput. 53 (2015) 331–344.
[78] S. Forsen, Graphical rules for enzyme-catalyzed rate laws, Biochem. J. 187 (1980) 829–835.
[79] G.P. Zhou, M.H. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways, Biochem. J. 222 (1984) 169–176.
[80] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics, J. Biol. Chem. 264 (1989) 12074–12079.
[81] I.W. Althaus, J.J. Chou, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E, Biochemistry 32 (1993) 6548–6554.
[82] 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.
[83] W.Z. Lin, X. Xiao, Wenxiang: a web-server for drawing wenxiang diagrams, Nat. Sci. 3 (2011) 862–865.
[84] 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.
[85] T. Fawcett, ROC graphs: notes and practical considerations for researchers, Mach. Learn. 31 (2004) 1–38.
[86] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240.