OPEN ACCESS Freely available online

Immunome Research

ISSN: 1745-7580

Research Article

Recent Progresses for Computationally Identifying N^6-methyladenosine Sites in *Saccharomyces cerevisiae*

Kuo-Chen Chou*

Gordon Life Science Institute, Boston, MA 02478, USA

ABSTRACT

N^6-methyladenosine (m^{6}A) plays critical roles in a broad set of biological processes. Knowledge about the precise location of m^{6}A site in the transcriptome is vital for deciphering its biological functions. Although experimental techniques have made substantial contributions to identify m^{6}A methylations, they are still labor intensive, costly and time consuming. As good complements to experimental methods, in the past few years, a series of computational approaches have been proposed to identify m^{6}A sites in *Saccharomyces cerevisiae*. In order to facilitate researchers to select appropriate methods for identifying m^{6}A sites, it is necessary to give a comprehensive review and comparison on existing computational methods. In this review, we summarized the current progresses in computational prediction of m^{6}A sites and also assessed the performance of computational methods for identifying m^{6}A sites on an independent dataset. Finally, challenges and future directions of computationally identifying m^{6}A sites were presented as well. Taken together, we anticipate that this review will provide an important guide for future computational analysis of m^{6}A and other RNA modifications.

Keywords: Post transcription modification; N^6-methyladenosine; Epitranscriptome; Machine learning method; 5-step rules

INTRODUCTION

Among the ~150 kinds of known RNA modifications, the N^6-methyladenosine (m^{6}A) is the most prevalent internal mRNA/lncRNA modifications, which occurs on the sixth nitrogen atom of adenine. As a reversible and dynamic post-transcriptional modification, the formation of m^{6}A is installed by a multicomponent methyltransferase complex including METTL3, METTL14 and WTAP, while its demethylation is regulated by demethylases FTO and ALKBH5. The biological functions of dynamic m^{6}A modification is regulated by m^{6}A readers, such as heterogeneous nuclear ribonucleoprotein C (HNRNPC), YTH Domain Family proteins 1, 2 and 3 (YTHDF1, YTHDF2, YTHDF3) and YTHDC1, etc. [1,2-4].

Since discovered in 1970s, m^{6}A has been observed in all three kingdoms of life. With the intensive researches on m^{6}A methylation in recent years, its functions have been uncovered gradually. It has been found that m^{6}A is associated with a broad set of fundamental cellular processes, such as RNA localization and degradation, RNA splicing, circadian rhythm, cell differentiation and reprogramming, immune tolerance and even the occurrence of diseases. However, few of them are currently understood in mechanistic detail. Identifying the precise location of m^{6}A site in transcriptomes will be of a great help to investigate its biological mechanisms and functions.

With the development of next-generation sequencing technology, the MeRIP-Seq and m^{6}A-seq high-throughput methods have been developed to identify m^{6}A sites in *Saccharomyces cerevisiae*, *Homo sapiens*, and *Mus musculus*. However, the resolution of these techniques is low and couldn't identify the exact methylated adenosines. Recently, the miCLIP technique was proposed, which provided the single-nucleotide resolution m^{6}A profile of the human transcriptome. Based on these experimental data, several informative databases related with m^{6}A modifications have been built. Taken together, these experiments promote the progress of researches on m^{6}A modifications. However, experimental techniques are still labor-intensive and expensive for transcriptome-

Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, Boston, MA 02478, USA, E-mail: kcchou@gordonlifescience.org

Received: October 26, 2020; **Accepted:** November 09, 2020; **Published:** November 16, 2020

Citation: Chou KC (2020) Recent Progresses for Computationally Identifying N^6-methyladenosine Sites in *Saccharomyces cerevisiae*. Immunome Res. 16:181.

wide detection of m^{6}A. Therefore, it is an urgent task to develop effective and low-cost approaches to automatically identify m^{6}A sites. As excellent complements to experimental techniques, computational methods are in high demand to accurately detect m^{6}A sites.

MATERIALS AND METHODS

In 2013, Schwartz proposed the first computational model to predict the m^{6}A site in the *S. cerevisiae* transcriptome, whose features include relative position in gene, nucleotide composition and predicted secondary structures. Although no public web server or software package was provided for this method, Schwartz's pioneer work provides a new strategy for identifying m^{6}A site. Since then, the scientific community witnessed an unprecedented amount of studies considering the application of machine learning method to identify m^{6}A sites. For example, a series of machine learning based methods, such as m6Apred, iRNA-Methyl, SRAMP, pRNAm-PC, RAM-ESVM, RAM-NPPS and RNA-MethylPred have been proposed. All these prediction methods were developed in principle by following the guidelines of the Chou's 5-step rule [5] as done in a series of powerful predictors developed recently for genome or proteome analyses.

Accordingly, these methods also share the advantages: (1) clearer in logic development, (2) more transparent in operation, and (3) more useful in practical application.

To provide the readership with a clear landscape about the recent developments in this important area, in this comprehensive review we are to elaborate their details in observing the Chou's 5-step rule [5]: (1) how to construct or select a valid benchmark dataset to train and test the predictor; (2) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) how to establish a user-friendly web-server for the predictor that is accessible to the public. Moreover, to facilitate users to select appropriate method according to their need, a comparison of existing methods in identifying m^{6}A sites is to be performed based on an independent dataset. Finally, the challenges and future perspectives for identifying m^{6}A sites are to be discussed.

Benchmark dataset for predicting m^6 A sites

Constructing a valid and reliable benchmark dataset is the critical step to train a computational model with high effectiveness [6,7]. In literature, the benchmark dataset usually consists of a training dataset and a testing dataset: the former is constructed for the purpose of training a proposed model, while the latter for the purpose of testing it. As pointed out by a comprehensive review [8], however, there is no need to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it is tested by the jackknife or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests.

For investigating the m^{6}A sites, the benchmark datasets were constructed as follows. The sequence segment surrounding the m^{6}A site contains the underlying discrimination information, whose size can be determined with the aid of sliding window scheme. The window length is usually set to $2n+1$, whose central element is the experimentally confirmed m^{6}A site, with n flanking nucleotides on both sides of the methylated adenosine. However, there is no uniform standard to set the window size. The determination of n is always associated with features extraction, prediction method and cross-validation performance.

In 2015, the first publicly available benchmark dataset (called S_1 here) for the prediction of m^{6}A sites was built by Chen. The positive samples in dataset S_1 are 832 m^{6}A sites with distances to the detected m^{6}A-seq peaks less than 10 bp, which were extracted from the 1,307 experimentally confirmed m^{6}A sites [53]. The negative samples in dataset S_1 are the 832 non-m^{6}A sites randomly selected from the 33,280 non-methylated adenines. Each sample in the dataset S_1 is 21-nt long with the m^{6}A sites or non-m^{6}A site in the center.

In some cases, the m^{6}A site locates at the beginning or end of the sequence, which results in that the extracted sequence fragments size is shorter than the given window size. Two strategies are often used to generate fixed window length. The first one is to fill the blank by using the dummy 'X' nucleotide that don't represent any real nucleotide. The second one is to fill the blank using the mirror image method. If the missing nucleotides locate at the beginning (i.e. upstream of the m^{6}A site), they will be filled by using their mirror images locate at downstream of the m^{6}A site, and vice versa (Figure 1). The second approach has been used to construct the benchmark dataset for the prediction of m^{6}A sites in *S. cerevisiae*.

In 2016, Chen built another benchmark dataset (called S_2 here) using the mirror image method, which includes 1,307 m^{6}A site containing sequences (positive samples) and the equal number of non-m^{6}A containing sequences (negative samples). In dataset S_2, all the experimentally confirmed m^{6}A sites in *S. cerevisiae* were included. The sequences in this dataset are 51-nt long with the sequence similarity less than 85%, with the m^{6}A site or non-m^{6}A site in the center. Since it has been built, nearly all the computational models for identifying m^{6}A site in *S. cerevisiae* were trained and tested on the dataset S_2 [54].

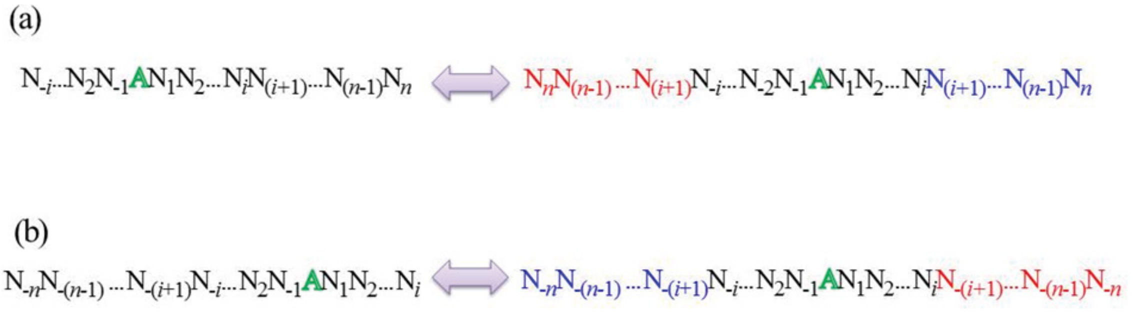
(a)
$N_{-i} \cdots N_2 N_{-1} A N_1 N_2 \cdots N_i N_{(i+1)} \cdots N_{(n-1)} N_n$
$N_n N_{(n-1)} \cdots N_{(i+1)} N_{-i} \cdots N_{-2} N_{-1} A N_1 N_2 \cdots N_i N_{(i+1)} \cdots N_{(n-1)} N_n$
(b)
$N_{-n} N_{-(n-1)} \cdots N_{-(i+1)} N_{-i} \cdots N_2 N_{-1} A N_1 N_2 \cdots N_i$
$N_{-n} N_{-(n-1)} \cdots N_{-(i+1)} N_{-i} \cdots N_2 N_{-1} A N_1 N_2 \cdots N_i N_{-(i+1)} \cdots N_{-(n-1)} N_{-n}$

Figure 1: A schematic illustration showing the mirror image for (a) upstream (b) downstream missing nucleotides, respectively. The real RNA segment is colored in blue and its mirror image is colored in red. The methylated A is highlighted in green.

Formulation of rna samples

The 2nd step of the 5-step rules [5] is about the formulation of biological samples. With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms (such as "Covariance Discriminant" or "CD" algorithm [9,10], "Nearest Neighbor" or "NN" algorithm [11,12], "Support Vector Machine" or "SVM" algorithm [13,14], and "Random Forest" or "RF" algorithm [15,16]) can only handle vectors as elaborated in a comprehensive review [17]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [18] or PseAAC [19] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (see example, [20-45] as well as a long list of references cited in [46]).

Because it has been widely and increasingly used, four powerful open access soft-wares, called 'PseAAC' [47], 'PseAAC-Builder' [48], 'propy' [49], and 'PseAAC-General' [50], were established: the former three are for generating various modes of Chou's special PseAAC [51]; while the 4th one for those of Chou's general PseAAC [5], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see equations..9-10 of [5]), "Gene Ontology" mode (see Equations 11-12 of [5]), and "Sequential Evolution" or "PSSM" mode (see equations13-14 of [5]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the idea of PseAAC was extended to PseKNC (Pseudo K-tuple Nucleotide Composition) to generate various feature vectors for DNA/RNA sequences [52] that have proved very successful as well [25,41,43,44,53-60]. Particularly, recently a very powerful web-server called 'Pse-in-One' [61] and its updated version 'Pse-in-One2.0' [62] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the users' need or their own definition." According to the concept of pseudo components, any RNA sequence sample can be formulated as [53].

$$\mathbf{R} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_u & \cdots & \phi_Z \end{bmatrix}^{\mathbf{T}} \tag{1}$$

where

$$\phi_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \le u \le 4^k) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w \sum_{j=1}^{\lambda} \theta_j} & \left(4^k < u \le 4^k + \lambda\right) \end{cases} \tag{2}$$

In Eq.2, f_u (u = 1,2,..., 4^k) is the normalized occurrence frequency of the u-th non-overlapping k-tuple nucleotide in the RNA sequence. λ is the number of the total pseudo components used to reflect the long-range or global sequence effect, and w is the weight factor. θ is the j-th tier correlation factor that reflects the sequence order correlation between all the j-th most contiguous k-tuple nucleotide along a L-nt long RNA sequence as formulated by

$$\theta_j = \frac{1}{L-k-j+1} \sum_{i=1}^{L-k-j+1} \mathrm{C}_{i,i+j} \qquad (j = 1, 2, \cdots, \lambda; \lambda < L-k) \tag{3}$$

where $\mathrm{C}_{i,i+j}$ is the correlation function and is defined by

$$\mathrm{C}_{i,i+j} = \frac{1}{\mu} \sum_{g=1}^{\mu} \left[P_g\left(\mathrm{R}_i \mathrm{R}_{i+1} \ldots \mathrm{R}_{i+k-1}\right) - P_g\left(\mathrm{R}_{i+j} \mathrm{R}_{i+j+1} \ldots \mathrm{R}_{i+j+k-1}\right) \right]^2 \tag{4}$$

where μ is the number of RNA physicochemical properties considered, P_g ($R_i R_{i+1}...R_{i+k-1}$) is the numerical value of the g-th (g=1, 2, 3, ..., u) RNA local structural property for the k-tuple nucleotide $\mathrm{R}_i\mathrm{R}_{i+1}...\mathrm{R}_{i+k-1}$ at position i and P_g ($R_{i+j} R_{i+j+1}...R_{i+j+k-1}$) the corresponding value for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+j+1}...\mathrm{R}_{i+j+k-1}$ at position $i + j$. The details about PseKNC can be found in our recent review article [53].

Algorithm or operation engine

The 3rd step of the 5-step rules [5] is about the operation engine. The two commonly used machine learning algorithms for identifying m6A sites are support vector machine (SVM) and random forest (RF), which were briefly introduced as following.

RESULTS AND DISCUSSION

Support vector machine (SVM)

SVM is a powerful and popular method for pattern recognition, which has been widely used in the realm of bioinformatics especially very effectively in a series of recent genome analyses (see example [63-65]). Its basic idea is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. Owing to its effectiveness and speed in training process, the radial basis kernel function (RBF) of SVM was often used to obtain the classification hyperplane. The regularization parameter C and kernel parameter γ of the SVM operation engine can be optimized in the following ranges [2^{-5}, 2^{15}] and [2^{-15}, 2^{-5}] with the steps of 2 and 2^{-1}, respectively. For a brief formulation of SVM and how it works, see the papers [66,67]. For more details about SVM, see a monograph [68].

Random forest (RF)

RF is an ensemble of a large number of decision trees. Each tree in the ensemble is trained on a subset of training instances and gives a classification result. The three parameters of RF, namely the number of trees, the number of features randomly selected, and the minimum number of samples required to split an internal node (nsplit) can be determined by using the grid search scheme. The predictive results of RF are based on the ensemble of those decision trees. Since proposed by Breiman in 2001 [69], owing to its advantages in dealing with high-dimensional data, RF has been used in many areas of bioinformatics (see example, [15,16,56,70-83]).

Performance evaluation

The 4th guideline of the 5-step rules [5] is about how to validate the proposed model. To address this, two issues are need to considered. One is what kind of metrics should be used to measure the scores, and the other is what test methods should be adopted to count the scores.

A set of intuitive metrics

The performance of the computational methods are usually evaluated using the following four metrics [84]: (1) overall accuracy or Acc, (2) Mathew's correlation coefficient or MCC, (3) sensitivity or Sn, and (4) specificity or Sp. However, the conventional metrics copied from math books are hard to be understood by most experimental scientists due to lacking intuitiveness; especially for the MCC, which is very important to indicate the stability of a predictor. Fortunately, using the symbols introduced by Chou [85] in studying signal peptide cleavage sites, a set of four intuitive metrics were derived [14,86], as given below

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_{-}^{+}}{N^{+}} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_{+}^{-}}{N^{-}} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \ \Lambda = 1 - \dfrac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_{-}^{+}}{N^{+}} + \dfrac{N_{+}^{-}}{N^{-}} \right)}{\sqrt{\left(1 + \dfrac{N_{+}^{-} - N_{-}^{+}}{N^{+}} \right)\left(1 + \dfrac{N_{-}^{+} - N_{+}^{-}}{N^{-}} \right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (5)$$

where N^{+} represents the total number of positive samples investigated, while N_{-}^{+} is the number of positive samples incorrectly predicted to be of negative one; N^{-} the total number of negative samples investigated, while N_{+}^{-} the number of the negative samples incorrectly predicted to be of positive one. The set of intuitive metrics have been concurred and applauded by a series of recent publications (see example, [14,16,57-59,74,82,83,87-100,101-114]). It is instructive point out, however, either the conventional metrics [84] taken from math books or the intuitive metrics of Equation 5 are valid only for single label systems (where each of the constituent samples belong to one, and only one, attribute or class); for the multi-label systems (where a sample may simultaneously belong to several different attributes or classes) whose existence has become more frequent in system biology [6,7,29,115-134], system medicine [135,136] and biomedicine [78,137], a completely different set of metrics as defined in [138] is absolutely needed.

Jackknife test

In statistical prediction, the following three cross-validation methods are often used to evaluate the performance of a predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [139]. Among them, however, the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, as elucidated in [5] and demonstrated by Equations 28-32 therein. Therefore, the jackknife test has been increasingly recognized and widely adopted by investigators to test the power of various prediction methods (see example, [140-146]). In view of this, the jackknife test was also adopted to evaluate the computational methods in identifying m^{6}A sites.

Web servers for detecting m^6 A sites

The last but not least important step of the Chou's 5-step rules [5] is about the web-server establishment. As pointed out in [147] and demonstrated in a series of recent publications (see example [55, 57-59,81,89,97,102,104,121-127,135, 48-160]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science [17], driving medicinal chemistry into an unprecedented revolution [46].

Computational methods for detecting m^6 A sites

Over the past several years, nine different computational methods were proposed to identify m^{6}A sites in the *S. cerevisiae* transcriptome. For clarity, their names and web server addresses (if available) are listed in Table 1 according to the chronological order. Show in Figure 2 is the corresponding flowchart.

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics, (see example [161-174], particularly in enzyme kinetics and protein folding rates [169, 175-177] as well as low-frequency internal motion [178,183].

Table 1: List of computational methods for identifying m^{6}A sites in *S. cerevisiae.*

Methods	Sn (%)	Sp (%)	Acc (%)	MCC
iRNA-Methyl	19.25	80.75	50	0
RAM-ESVM	18.83	68.62	43.72	-0.14
RNA-MethylPred	27.2	74.06	50.63	0.01
RAM-NPPS	28.87	71.55	50.21	0
DeepM6APred	48.54	70.29	59.41	0.19

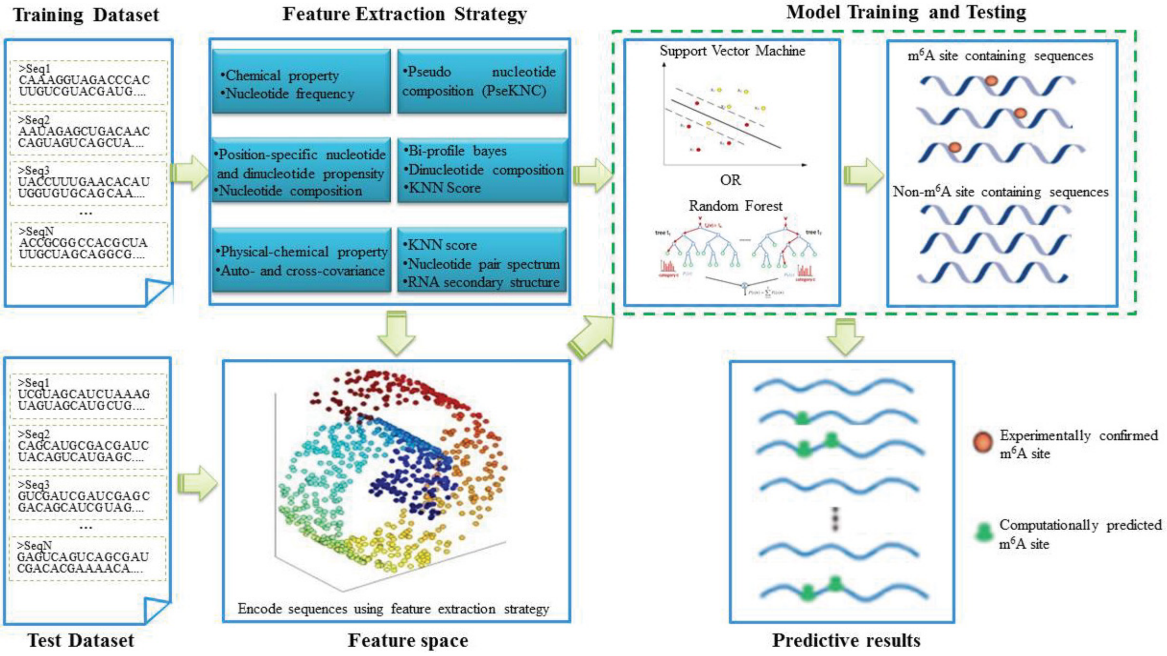

Figure 2: The general framework of computational method for identifying m^{6}A sites. The widely used feature extraction strategies and machine learning classifiers were shown in this figure.

Below, we are to make a comparison among the nine different prediction methods via their flowcharts as well.

m6Apred

Inspired by Schwartz's pioneer work, a support vector machine (SVM) based method called m6Apred was proposed by Chen, which encodes RNA sequences by using both accumulated nucleotide frequency and nucleotide chemical properties (i.e. chemical structure, chemical binding and chemical functionality).

Compared with the classic nucleotide composition, the accumulated nucleotide frequency includes not only the nucleotide frequency information, but also the distribution of each nucleotide in the RNA sequence. The three kinds of nucleotide chemical properties have different impacts on RNA's low-frequency internal motion and its biological function.

Accordingly, each nucleotide in the sequence was represented by a 4-dimenstional vector, in which the first element is the accumulated nucleotide frequency and the remaining three elements correspond to the nucleotide chemical properties. For the sequence with a length of *L* (where *L*=21), it can be represented by a 4*L*-dimenstional vector and used as the input of SVM. The proposed m6Apred obtained a satisfactory performance for identifying m^{6}A site in the *S. cerevisiae* transcriptome based on dataset $\mathbb{S}_1$.

iRNA-Methyl

It was found that the formation of m^{6}A methylation is affected by RNA secondary structure that is highly related with the physicochemical properties of dinucleotide. In view of this, a predictor called "iRNA-Methyl" was proposed by formulating RNA sequences with the pseudo nucleotide composition (PseKNC). By using PseDNC (i.e. k=2 in Equation 2), three physicochemical properties, namely enthalpy, entropy, and free energy that can quantify the RNA secondary structures were used to calculate the long-range sequence order effects using the following formula:

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 16) \\[2ex] \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < u \le 16+\lambda) \end{cases} \tag{6}$$

where f_u (u = 1,2,..., 16) is the normalized occurrence frequency of the u-th non-overlapping dinucleotide in the RNA sequence, and

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} \mathrm{C}_{i,i+j} \qquad (j = 1,2,\cdots,\lambda;\lambda < L) \tag{7}$$

where θ_j is called the *j*-tier correlation factor that reflects the sequence order correlation between all the *j*-th most contiguous dinucleotide, the coupling factor $\mathrm{C}_{i,i+j}$ is given by

$$\mathrm{C}_{i,i+j} = \frac{1}{3}\sum_{g=1}^{3}\left[P_g(\mathrm{D}_i) - P_g(\mathrm{D}_{i+j})\right]^2 \tag{8}$$

where $P_g(D_i)$ (g=1, 2, 3) is the normalized value of the above mentioned three RNA physicochemical properties for the *i*-th dinucleotide D_i. By using the 10-fold cross validation test, the optimal values for the parameters λ and *w* of PseKNC were obtained (i.e. λ=6 and *w*=0.9). Accordingly, the samples in dataset $\mathbb{S}_2$ were transferred into a 22-dimensional vector in the iRNA-Methyl method. It was found that iRNA-Methyl obtained an accuracy of 65.59% for identifying m^{6}A sites in the rigorous jackknife test. For the convenience of experimental scientists, a web-server for iRNA-Methyl has been established at http://lin-group.cn/server/iRNA-Methyl.

pRNAm-PC

In 2016, in order to improve the accuracy of m^{6}A site identification, Liu proposed the pRNAm-PC method, in which the RNA sequences in dataset $\mathbb{S}_2$ were encoded by using a vector, whose components were derived from a physical-chemical matrix via the auto-covariance and cross-covariance transformations [55].

Based on the dinucleotide physicochemical properties, a 10×50 dimensional physicochemical property matrix (PC) was defined as following,

$$\mathrm{PC} = \begin{bmatrix} \mathrm{P_1(N_1N_2)} & \mathrm{P_1(N_2N_3)} & \ldots & \mathrm{P_1(N_iN_{i+1})} & \ldots & \mathrm{P_1(N_{50}N_{51})} \\ \mathrm{P_2(N_1N_2)} & \mathrm{P_2(N_2N_3)} & \ldots & \mathrm{P_2(N_iN_{i+1})} & \ldots & \mathrm{P_2(N_{50}N_{51})} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{P_j(N_1N_2)} & \mathrm{P_j(N_1N_2)} & \ldots & \mathrm{P_j(N_iN_{i+1})} & \ldots & \mathrm{P_j(N_{50}N_{51})} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{P_9(N_1N_2)} & \mathrm{P_9(N_2N_3)} & \ldots & \mathrm{P_9(N_iN_{i+1})} & \ldots & \mathrm{P_9(N_{50}N_{51})} \\ \mathrm{P_{10}(N_1N_2)} & \mathrm{P_{10}(N_2N_3)} & \ldots & \mathrm{P_{10}(N_iN_{i+1})} & \ldots & \mathrm{P_{10}(N_{50}N_{51})} \end{bmatrix} \tag{9}$$

where $P_j(N_iN_{i+1})$ is the *j*-th (*j*=1, 2, ..., 10) physicochemical properties value for the dinucleotide N_iN_{i+1} (namely AA, AC, AG, AU, CA, ..., or UU) in the RNA sequence. In the pRNAm-PC method, 10 dinucleotide physicochemical properties (i.e. P_1: rise, P_2: roll, P_3: shift, P_4: slide, P_5: tilt, P_6: twist, P_7: enthalpy, P_8: entropy, P_9: stack energy, and P_{10}: free energy) were used.

In order to reflect the correlation of the same and different physicochemical property between two subsequences separated by λ dinucleotides, the auto-covariance (AC) and cross-covariance (CC) method were used to transform the physicochemical property matrix into a length-fixed feature vector and were defined as following.

$$\mathrm{AC}(\mathrm{j},\lambda) = \frac{\sum_{i=1}^{50-\lambda}[\mathrm{P}_j(\mathrm{N}_i\mathrm{N}_{i+1}) - \bar{\mathrm{P}}_j][\mathrm{P}_j(\mathrm{N}_{i+\lambda}\mathrm{N}_{i+1+\lambda}) - \bar{\mathrm{P}}_j]}{50-\lambda} \tag{10}$$

$$\mathrm{CC}(\mathrm{j},\mathrm{k},\lambda) = \frac{\sum_{i=1}^{50-\lambda}[\mathrm{P}_j(\mathrm{N}_i\mathrm{N}_{i+1}) - \bar{\mathrm{P}}_j][\mathrm{P}_k(\mathrm{N}_{i+\lambda}\mathrm{N}_{i+1+\lambda}) - \bar{\mathrm{P}}_k]}{50-\lambda} \quad (j \ne k) \tag{11}$$

By preliminary tests, they found that the best value for λ is 4. Therefore, the RNA sequences in dataset $\mathbb{S}_2$ were encoded by a 400-dimensional vector, of which the 40 elements were deduced from the auto-covariance and the 360 elements from the cross-covariance. Based on this kind of feature, the pRNAm-PC was built and yielded an accuracy of 69.74% for identifying the m^{6}A sites in dataset $\mathbb{S}_2$ in the jackknife test, which is ~5% higher than that of iRNA-Methyl. However, the feature dimension of pRNAm-PC was nearly 26 times larger than that of iRNA-Methyl. Moreover, the contributions and biological meanings of the above mentioned 10 physicochemical properties for identifying m^{6}A sites were not described at all.

SRAMP

Subsequently, by combining multiple features, Zhou and his colleagues established a random forest based computational predictor, called SRAMP, which is available at http://www.cuilab.cn/sramp/. In order to capture more sequence-derived features, the positional nucleotide sequence pattern, K-nearest neighbor information, the position independent nucleotide pair spectrum, and the predicted RNA secondary structure were used to encode the RNA sequences.

For the positional nucleotide sequence pattern, the nucleotide (A,C,G or U) at each position were represented by the binary vector of (1,0,0,0), (0,1,0,0), (0,0,1,0), or (0,0,0,1). For a 2*n*+1 long sequence segment, a 4×(2*n*+1) dimensional vector can be obtained.

In order to measure the extent of how much the flanking window of one query sample resembles those of other m^{6}A sites, the K-nearest neighbor information was introduced. Firstly, the flanking window of the query sample was compared with all samples in the training dataset and obtained a pair-wise similarity score,

$$Score = \sum_{i=1}^{2n+1} \mathrm{NUC44}(q_i, r_i) \tag{12}$$

where q_i and r_i are the nucleotides at the *i*-th position of the flanking windows in the query sample and the training samples. 2*n* + 1 is the window size. The NUC44 is a common nucleotide similarity scoring matrix. And then, the fraction of positive samples in the top *K* most similar reference samples was taken as the KNN feature. In SRAMP, 30 K values were used (i.e., K=50, 100, 150, ..., 1500).

The sequence context was also reflected by calculating the frequencies of all possible *d*-spaced nucleotide pairs, which is defined as

$$Frequency(np_i) = \frac{C(np_i)}{2n-d-1} \tag{13}$$

where C(np_i) is the number of np_i inside a flanking window with a size of 2*n*, *d* is the space between two nucleotides, and ranged from 0 to 3.

As indicated in their work, the hairpin loop, multiple loop, interior loop, paired and bulged loop from the RNA secondary structure were also used to represent RNA sequences, which were encoded as the binary vectors, namely (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1) and (0,0,0,0,0), respectively. Accordingly, in the structure space, a flanking window with a size of 2*n* will be converted into a 2*n*×5-dimensional vector.

For each kind of these features, a random forest classifier was built. The final prediction result was the combination of them by using the weighted summing scheme. As indicated by, SRAMP yielded comparable accuracy for identifying m^{6}A sites in the *S. cerevisiae* transcriptome to that of iRNA-Methyl. As a big plus, SRAMP can not only identify m^{6}A sites in *S. cerevisiae*, but also much more effective for identifying mammalian m^{6}A sites.

M6A-HPCS

Later on, with the aim of finding out which physicochemical properties making great contributions for identifying m^{6}A sites, Zhang proposed a heuristic nucleotide physicochemical property selection algorithm, called M6A-HPCS, to identify m^{6}A sites in the *S. cerevisiae* transcriptome [99]. The M6A-HPCS method is based on the iRNA-Methyl and pRNAm-PC methods. However, rather than directly using the physicochemical properties, the relative gain and direct gain methods were used to measure the significance of each of the 23 dinucleotide physicochemical properties for identifying the m^{6}A sites. And then, a heuristic algorithm is employed to select the optimal physicochemical properties for the PseKNC (used in iRNA-Methyl) and auto-covariance and cross-covariance (used in pRNAm-PC) encoding schemes, respectively.

For the PseKNC and auto-covariance and cross-covariance encoding scheme, 5 and 13 out of the 23 dinucleotide physicochemical properties were selected out as their optimal candidates to represent the RNA sequences in dataset S_2, respectively. In the rigorous jackknife test, the accuracies of 67.33% and 72.38% were obtained for identifying m^{6}A sites for both encoding schemes, respectively. Although its predictive accuracy is higher than those of iRNA-Methyl and pRNAm-PC, the shortcomings for M6A-HPCS still exist in the following aspects. First, the biological meanings of using the optimal dinucleotide physicochemical properties are not described at all. Second, although a web-server was developed for M6A-HPCS at http://csbio.njust.edu.cn/bioinf/M6A-HPCS, it couldn't be accessed anymore.

RNA-MethylPred

To further improve the accuracy of m^{6}A site identification, Jia proposed a new computational method called RNA-MethylPred. In this method, three kinds of feature extraction strategies were used to represent the RNA sequences in dataset S_2 [16].

Bi-profile bayes vector (V) was employed to reflect the posterior probability of positive and negative samples.

$$V= [p_1, p_2,..., p_n, p_{n+1},..., p_{2n}] \tag{14}$$

where the first *n* components denotes the posterior probability of each nucleotide at the *i*-th position in the positive samples, the remaining *n* components denotes the posterior probability of each nucleotide at the *i*-th position in the negative samples. *n* is equal to the length of the RNA sequence (i.e. *n*=51).

Two forms of dinucleotide composition were defined to reflect sequence order information.

$$P_{ab} = \frac{N_{ab}}{N_{a.}} \tag{15}$$

$$P'_{ab} = \frac{N_{ab}}{n-1} \tag{16}$$

where N_{ab} is the number of neighboring dinucleotide (*a*, *b* can be nucleotide A, C, G or U), *a*• indicates any adjoining dinucleotides that starting with *a*.

K nearest neighbor (KNN) scores were used to measure whether the local sequence similarity. To this end, similarity score S(A,B) between two sequence fragments A and B was defined as

$$S(A,B) = \sum_{1\le i\le 51} Score(A[i], B[i]) \tag{17}$$

A[i] indicates the nucleotide at the *i*-th position in sequence segment A, and the score of two nucleotides was defined as

$$Score = \begin{cases} +2 & when\,two\,nucleotides\,matched \\ -1 & when\,two\,nucleotides\,mismatched \end{cases} \tag{18}$$

Based on Equations. (11) and (12), the KNN score was achieved by calculating the percentage of the positive neighbors in its KNNs. In RNA-MethylPred, the 20 considered Ks were 10, 20, ..., 200.

Finally, these features were combined together and used as the input of SVM to perform the prediction. In the jackknife test, the RNA-MethylPred obtained an accuracy of 76.51% for identifying m^{6}A sites in the *S. cerevisiae* transcriptome. Rather than building a web-server, the authors provided a MATLBA package for the RNA-MethylPred method.

RAM-ESVM

As introduced above, various features and predictors have been proposed for identifying m^{6}A sites. However, their performances are still not satisfactory. In 2017, Chen developed an ensemble classifier called RAM-ESVM, which combines three basic classifiers based on different features including PseKNC, motif features, and optimized K-mer [57]. The first two classifiers (SVM-PseKNC and SVM-motif) were built based on SVM by using PseKNC and motif features as the inputs, respectively. The third one is also a SVM based classifier and its input features are optimized gapped *k*-mers, which is achieved by using the GkmSVM software. The three basic classifiers vote for the final result based on the voting score.

$$V_i = \sum_{k=1}^{3} f\left(pre\left(C_k\right), Class_i\right) \qquad (i=1,2; k=1,2,3) \tag{19}$$

where V_i is the voting score for the RNA sample belonging to the class$_i$ (i=1: m6A sites; i=2: non- m6A sites), and

$$f\left(pre\left(C_k\right), Class_i\right) = \begin{cases} 1: & if\ pre\left(C_k\right) \in Class_i \\ 0: & otherwise \end{cases} \tag{20}$$

The final prediction is determined by the argument that maximizes the voting score V_i,

$$\text{Sgn}(i) = arg\ max_i\{V_i\} \tag{21}$$

In the jackknife test, the RAM-ESVM produced an accuracy of 78.35% for identifying m^{6}A sites in the *S. cerevisiae* transcriptome. The RAM-ESVM can be freely accessed at http://server.malab.cn/RAM-ESVM/.

RAM-NPPS

In 2017, another m^{6}A site predictor, called RAM-NPPS, was proposed by Xing, which is based on multi-interval nucleotide pair position specificity (NPPS).

For a given RNA sequence segment P, it can be represented by P=P$^+$-P$^-$. P$^+$ and P$^-$ can be formulated as

$$P^{\xi} = p_1^{\xi} p_2^{\xi} \ldots p_k^{\xi} \ldots p_{51}^{\xi} \qquad (\xi \text{ is} + or -) \tag{22}$$

To obtain p_k^{ξ}, single nucleotide position matrix T_s^{ξ} and dinucleotide position matrix T_d^{ξ} are defined as following

$$T_s^{\xi} = \begin{bmatrix} f_{A,1}^{\xi} & f_{A,2}^{\xi} & \cdots & f_{A,51}^{\xi} \\ f_{C,1}^{\xi} & f_{C,2}^{\xi} & \cdots & f_{C,51}^{\xi} \\ f_{G,1}^{\xi} & f_{G,2}^{\xi} & \cdots & f_{G,51}^{\xi} \\ f_{U,1}^{\xi} & f_{U,2}^{\xi} & \cdots & f_{U,51}^{\xi} \end{bmatrix} \tag{23}$$

$$T_d^{\xi} = \begin{bmatrix} f_{AA,1}^{\xi} & f_{AA,2}^{\xi} & \cdots & f_{AA,50}^{\xi} \\ f_{AC,1}^{\xi} & f_{AC,2}^{\xi} & \cdots & f_{AC,50}^{\xi} \\ \vdots & \vdots & \cdots & \vdots \\ f_{UU,1}^{\xi} & f_{UU,2}^{\xi} & \cdots & f_{UU,50}^{\xi} \end{bmatrix} \tag{24}$$

The elements in the two matrices indicate the occurrence probability of the nucleotide in each position and the occurrence probability of the nucleotide pair at position i and i+η, respectively.

Suppose the dinucleotide between the i-th nucleotide and (i+η)-th nucleotide is 'ab', p_k^{ξ} can be calculated according to the conditional probability,

$$p_k^{\xi} = \frac{P(a \cap b)}{P(b)} \quad \frac{f_{ab,i}^{\xi}}{f_{b,i}^{\xi}} \tag{25}$$

Accordingly, the RNA sequence can be converted into the feature vector as described in Eq. 18. When the optimal interval value of the two nucleotides is set as η=5, the SVM based computational model RAM-NPPS was built, which is available at http://server.malab.cn/RAM-NPPS/. In the jackknife test, RAM-NPPS yielded an accuracy of 79.92% for identifying m^{6}A sites in the *S. cerevisiae* transcriptome.

DeepM6APred

More recently, Wei proposed a new method called DeepM6APred, which represented the RNA samples by using both the above mentioned NPPS feature and binary string encoding scheme [63]. Different from traditional methods, before directly using these features to make predictions, the deep-belief network was used to automatically learn meaningful feature representations from raw input sequences. Finally, an optimal feature set containing 429 features was obtained, based on which a predictive accuracy of 80.50% was obtained for identifying m^{6}A sites in the *S. cerevisiae* transcriptome. To the best of our knowledge, this is the best accuracy for identifying m^{6}A sites in the *S. cerevisiae* transcriptome till now. The DeepM6APred can be accessed at http://server.malab.cn/DeepM6APred.

Comparison of various prediction methods

In this section, we performed a comparison on existing methods for identifying m^{6}A sites in the *S. cerevisiae* transcriptome. Since SRAMP is a mammalian specific predictor and m6Apred is trained based on dataset $\mathbb{S}_1$, for a fair comparison, they were not considered here. The predictive accuracies of the other 7 methods for identifying m^{6}A sites based on the benchmark dataset $\mathbb{S}_2$ were shown in Figure 3. It was found the performance of DeepM6APred ranks the top.

To further demonstrate the generalization ability of these methods, an independent dataset was built, which includes 239 m^{6}A site containing sequences obtained from the RMBase, and the same number of non-m^{6}A site containing sequences. All these sequences are 51 nt and independent from the samples in the dataset $\mathbb{S}_2$, which are available at https://github.com/chenweiimu/m6a.

It should be point out that the web-server of M6A-HPCS is not accessible anymore as indicated in its homepage, and the pRNAm-PC could not make predictions for these independent sequences. Therefore, the comparisons were performed among the remaining methods (i.e. iRNA-Methyl, RAM-ESVM, RNA-MethylPred, RAM-NPPS, and DeepM6APred). Their predictive results for identifying m^{6}A sites in the independent dataset were reported in Table 2. It was found that the Sn, Acc and MCC of DeepM6APred are much higher than the other four methods. Although iRNA-Methyl obtained a high Sp, it has lower Sn, Acc and MCC. Thus, we can draw a conclusion that the performance of DeepM6APred is the best, while the performance of iRNA-Methyl is comparable to RNA-MethylPred and RAM-NPPS.

As demonstrated by many previous studies on a series of important

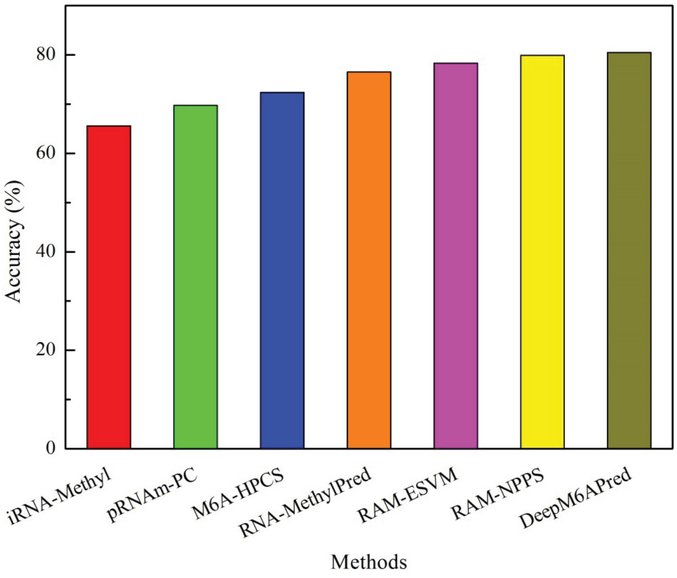

Figure 3: The performance of different methods for identifying m^{6}A sites in the benchmark dataset $\mathbb{S}_2$.

biological topics (see example [161,162,166,168,170,173,180-182]), using image or graphic approaches to study biological systems can provide intuitive insights for helping analyze complicated relations therein, in view of this, the accurately predicted m^6A sites by the different methods were presented in Figure 4. As we can see from the figure, of the 239 m^6A site containing sequences, 116 were correctly identified by DeepM6APred, 69 by RAM-NPPS, 65 by RNA-MethylPred, 46 by iRNA-Methyl, and 45 by RAM-ESVM. These results indicate that for users who are interested in identifying m^6A sites in the *S. cerevisiae* transcriptome, the DeepM6APred predictor should be their first choice, and the other predictors, namely RAM-NPPS, RNA-MethylPred, iRNA-Methyl and RAM-ESVM, may used as complementary tools in this regard.

Key Points

It is an urgent task to develop effective computational approaches for detecting m^6A and other RNA modifications in the transcriptome.

Table 2: Performance comparisons of different methods for identifying m^6A sites based on independent dataset. See Eq.5 for the definition of metrics below.

Methods	Web server address	Reference
Schwartz's method (2013)	Not available	{Schwartz, 2013 #29}
m6Apred (2015)	http://lin-group.cn/server/m6Apred	{Chen, 2015 #30}
iRNA-Methyl (2015)	http://lin-group.cn/server/iRNA-Methyl	{Chen, 2015 #31}
SRAMP (2016)	http://www.cuilab.cn/sramp/	{Zhou, 2016 #32}
M6A-HPCS (2016)	http://csbio.njust.edu.cn/bioinf/M6A-HPCS/	{Zhang, 2016 #39}
pRNAm-PC (2016)	http://www.jci-bioinfo.cn/pRNAm-PC	{Liu, 2016 #33}
RNA-MethylPred (2016)	Not available	{Jia, 2016 #36}
RAM-ESVM (2017)	http://server.malab.cn/RAM-ESVM/	{Chen, 2017 #34}
RAM-NPPS (2017)	http://server.malab.cn/RAM-NPPS/	{Xing, 2017 #35}
DeepM6APred (2018)	http://server.malab.cn/DeepM6APred	{Wei, 2018 #40;Wei,2018 #40}

The computational approaches for identifying m^6A sites in the *S. cerevisiae* transcriptome are introduced and discussed.

To help biologists choose appropriate methods for identifying m^6A sites, a comprehensive comparison on existing methods was performed on the independent dataset.

The challenges and future directions for identifying RNA modifications were discussed.

Remarks and perspective

In this paper, we comprehensively reviewed the computational methods for identifying m^6A sites in the *S. cerevisiae* transcriptome and evaluated their performance based on the independent dataset. Although these methods obtained quite good results in identifying m^6A sites when tested by the benchmark dataset S_2 of Section 2, their exploited or extrapolative effectiveness in practical application [139] was not so ideal as reflected by the fact when tested by the independent dataset.

The poor performance of these methods on the independent dataset is due to the following reason. All these methods were trained based on dataset S_2, in which both positive and negative samples were obtained by selecting the sequences containing RGAC consensus motif. However, in most cases, the m^6A site may not locate in the RGAC consensus motif. Thus, the construction of the benchmark dataset in such a way precluded the generalization ability of these methods. In order to improve the performance and generalization ability of the computational methods for identifying m^6A sites, much more efforts should be made by considering the following aspects.

The performance is dependent on the benchmark dataset

Although several benchmark datasets have been established for training computational models for identifying m^6A sites, the challenges still exist in the construction of the benchmark dataset.

Compared with the positive samples, there is no uniform standard to collect negative samples (non-m^6A samples). The popular strategy of obtaining non-m^6A samples is to select the adenosines that are not experimentally annotated as being methylated. It indeed raises the possibility that the m^6A sites are not identified may serve as false negative samples. In addition, in the real case, the number of non-m^6A sites is significantly higher than that of m^6A sites. The

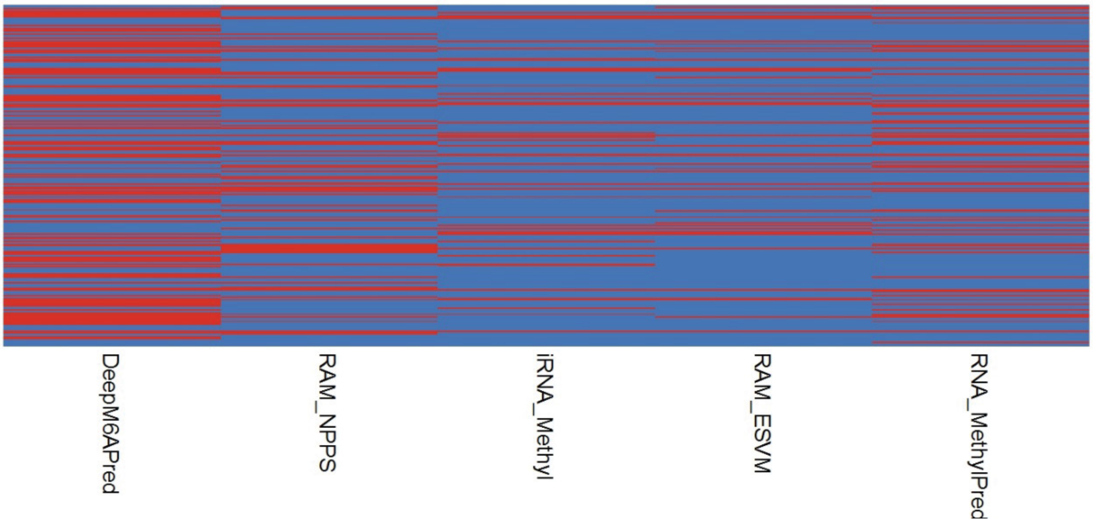

Figure 4: The detail predictive results of DeepM6APred, RAM-NPPS, iRNA-Methyl, RAM-ESVM and RNA-MethylPred based on the independent dataset. Each row is a sample in the independent dataset. The correctly identified m^6A site containing samples were highlighted in red, and the counterparts were in blue.

existing benchmark datasets are all balanced ones that contain roughly equal number of m6A samples and the randomly selected non-m6A sample. Such randomly sampling of non-m6A samples may lead to inadequate learning and the models trained on such a dataset would change when the selected non-m6A samples are different. To solve these challenges, more efforts can be made in the following aspects.

First, the one-sided selection (OSS) undersampling and synthetic minority oversampling technique (SMOTE) can be used to balance the non-m6A and m6A samples and minimize the influences of imbalance issue. The one-sided selection (OSS) undersampling employed the condensed nearest-neighbor to remove redundant negative samples that are far from the boundary of the class and the Tomek links to eliminate borderline samples and samples suffering from class label noise. By doing so, the number of non-m6A samples can be decreased. On the other hand, the SMOTE will resample the small class (m6A samples) by taking each small class example and introducing synthetic examples along the line segments joining it to the small class nearest neighbors. Accordingly, the positive and negative samples will be balanced.

The second way to deal with such an imbalance problem is to use ensemble techniques, which trains basic classifiers with different sampling data and combines their results to reduce the random sampling bias. The key step of this technique is to select meaningful negative samples to train basic classifiers.

Another strategy is to use cost-sensitive classifiers, such as XGboost (eXtreme Gradient Boosting), which can be trained with all the samples without selecting a subset of negative samples and prevent training model from over-fitting by defining different costs for the misclassified positive and negative samples.

Encode RNA sequences using effective schemes

Feature extraction strategy is another essential step to build computational models for identifying m6A sites. The performance of existing models for identifying m6A sites depends on how to accurately represent RNA sequences. The encoding schemes are based on the experiences and usually derived from the segments surrounding the m6A sites, such as pseudo nucleotide compositions, physicochemical properties, position specific nucleotide/dinucleotide composition, and so on. Although considerable progresses have been achieved, the following aspects should be considered for designing distinguishable feature descriptors in the future work.

Except for Schwartz's and Zhou's works, none of the other existing computational methods represented the RNA samples using RNA secondary structure information [55]. By regulating the interaction of methyltransferase complex with RNA sequences, RNA secondary structure is closely related to the formation of m6A. Therefore, it is necessary to integrate this kind of feature when constructing more powerful computational models for identifying m6A sites. To this end, the RNAfold tool in ViennaRNA package can be used to predict RNA secondary structure, whose output is dots (indicate unpaired nucleotides) and brackets (indicate paired nucleotides). If encode unpaired nucleotides using 0 and the paired one using 1, a given RNA sample will be transferred into a feature vector with its elements are 0 and 1.

Another shortcoming of existing methods is that existing computational methods directly use the entire features, which may lead to over-fitting problems, reduce the generalization capacity of the model and increase the computational time. In order to alleviate irrelevant features and overcome the above mentioned shortcoming, the feature selection techniques, such as minimal redundancy maximal relevance (mRMR), maximum relevancy maximum distance (MRMD), and analysis of variance (ANOVA), can be used to winnow out the optimal features.

Generalizability of existing computational approaches

Compared with the performance for identifying m6A sites in other species, the accuracy for identifying m6A sites in the *S. cerevisiae* transcriptome is still far from satisfactory. Therefore, new computational models are still required. Besides support vector machine and random forest, other machine learning methods such as Native Bayes, Logistic Regression, and K-nearest neighbor are all potential candidates to build new computational models for identifying m6A sites. With the development of convolutional neural network and deep learning, these advantaged approaches are also suggested to be used in developing computational models. In addition, since most of the existing methods are complementary to each other (Figure 4), it's wise to employ the ensemble classification techniques to develop computational models with high performance.

CONCLUSION

Besides m6A, the pseudouridine, N1-Methyladenosine (m1A), and 5-methylcytosine (m5C) are also frequently observed RNA modifications. However, both the computational models and experimental techniques couldn't simultaneously identify these different types of RNA modifications. To address such a challenge, more efforts should be made to develop a platform that can be used to simultaneously detect different types of RNA modifications.

ACKNOWLEDGEMENT

The authors thank Dr Cangzhi Jia for her assistance in running the RNA-MethylPred program.

FUNDING

This work was supported by the National Nature Scientific Foundation of China (31771471, 61772119), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244).

REFERENCES

1. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W, Chou KC. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels BioMed Res Internat. 2014;286419.
2. Fan GL, Zhang XY, Liu YL, Nang Y, Wang H. DSPMP: Discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns. J Comput Chem. 2015;36:2317-2327.
3. Pan L, Zhao W, Lai J, Ding D, Zhang Q, Yang X, et al. Sortase A-Generated Highly Potent Anti-CD20-MMAE Conjugates for Efficient Elimination of B-Lineage Lymphomas. Small. 2017;13
4. Oxenoid K, Dong YS, Cao C, Cui T, Sancak Y, Markhard AL, et al. Architecture of the Mitochondrial Calcium Uniporter. Natu. 2016;533:269-273.
5. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review, 5-steps rule). J Theor Biol. 2011;273:236-247.
6. Chou KC, Shen HB. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. Natu Proto. 2008;3:153-162.

7. Chou KC, Shen HB. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natu Sci. 2010;2:1090-1103.

8. Chou KC, Shen HB. Recent progresses in protein subcellular location prediction. Anal Biochem. 2007;370:1-16.

9. Chou KC, Elrod DW. Bioinformatical analysis of G-protein-coupled receptors. J Prot Res. 2002;1:429-433.

10. Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC. iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. PLoS ONE. 2012;7:47843.

11. Cai YD, Chou KC. Predicting subcellular localization of proteins in a hybridization space. Bioinform. 2004;1151-1156.

12. Chou KC, Cai YD. Prediction of protease types in a hybridization space. Biochem Biophys Res Comm. 2006;339:1015-1020.

13. Feng PM. Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem. 2013;442:118-125.

14. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nuclei Aci Res. 2013;41:68.

15. Lin WZ, Fang JA, Xiao X, Chou KC. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE. 2011;6:24756.

16. Jia J, Liu Z, Xiao X, Liu B, Chou KC. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J Theor Biol. 2016;394:223-230.

17. Chou KC. Impacts of bioinformatics to medicinal chemistry. Med Chem. 2015;11:218-234.

18. Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Genet. 2001;43:246-255.

19. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinform. 2005;21:10-19.

20. Chou KC, Cai YD. Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem. 90;2003:1250-1260.

21. Arif M, Hayat M, Jan Z. iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition. J Theor Biol. 2018;442:11-21.

22. Mei J, Zhao J. Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. Sci Rep. 2018;8:2359.

23. Mei J, Zhao J. Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features. J Theor Biol. 2018;427:147-153.

24. Krishnan MS. Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. J Theor Biol. 2018;445:62-74.

25. Zhang L, Kong L. iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components. J Theor Biol. 2018;441:1-8.

26. Zhang S, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J Theor Biol. 2018;437:239-250.

27. Butt AH, Rasool N, Khan YD. Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC. Mol biol rep. 2018;18:39-58.

28. Contreras-Torres E. Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC, J Theor Biol. 2018;454:139-145.

29. Javed V, Hayat M. Predicting subcellular localizations of multi-label proteins by incorporating the sequence features into Chou's PseAAC. Genom. 2018;17:793-821.

30. Ju Z, Wang SY. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. Gene. 2018;664:78-83.

31. Liang Y, Zhang S. Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback-Leibler divergence. J Theor Biol. 2018;454:22-29.

32. Mei J, Fu Y, Zhao J, Analysis and prediction of ion channel inhibitors by using feature selection and Chou's general pseudo amino acid composition. J Theor Biol. 2018;456:41-48.

33. Mousavizadegan M, Mohabatkar H. Computational prediction of antifungal peptides via Chou's PseAAC and SVM. J bioinform comput boil. 2018;1850016.

34. Qiu W, Li S, Cui X, Yu Z, Wang M, Du J, et al. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. J Theor Biol. 2018;450:86-103.

35. Rahman SM, Shatabda S, Saha S, Kaykobad M, Rahman MS. DPP-PseAAC: A DNA-binding Protein Prediction model using Chou's general PseAAC. J Theor Biol. 2018;452:22-34.

36. Sankari ES, Manimegalai DD. Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC. J Theor Biol. 2018;455:319-328.

37. Srivastava A, Kumar R, Kumar M. BlaPred: predicting and classifying beta-lactamase using a 3-tier prediction system via Chou's general PseAAC. J Theor Biol. 2018;457:29-36.

38. Zhang S, Liang Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. J Theor Biol. 2018;457:163-169.

39. Zhao W, Wang L, Zhang TX, Zhao ZN, Du PF. A brief review on software tools in generating Chou's pseudo-factor representations for all types of biological sequences. Prot Pept Lett. 2018;25:822-829.

40. Akbar S, Hayat M. iMethyl-STTNC: Identification of N(6)-methyladenosine sites by extending the Idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol. 2018;455:205-211.

41. Al Maruf MA, Shatabda S. iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components. Genom. 2018;18:63-82.

42. Pan Y, Wang S, Zhang Q, Lu Q, Su D, Zuo Y, et al. Analysis and prediction of animal toxins by various Chou's pseudo components and reduced amino acid compositions. J Theor Biol. 2019;462:221-229.

43. Tahir M, Hayat M, Khan SA. iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition. Mol gene genom. 2019;294:199-210.

44. Tahir M, Tayara H, Chong KT. iRNA-PseKNC(2methyl): Identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components. J Theor Biol. 2019;465:1-6.

45. Tian B, Wu X, Chen C, Qiu W, Ma Q, Yu B. Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. J Theor Biol. 2019;462:329-346.

46. Chou KC. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Cur Top Med Chem. 2017;17: 2337-2358.

47. Shen HB, Chou KC. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem. 2008; 373:386-388.

48. Du P, Wang X, Xu C, Gao Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions. Anal Biochem. 2012;425:117-119.

49. Cao DS, Xu QS, Liang YZ. Propy: a tool to generate various modes of Chou's PseAAC. Bioinform. 2013;29:960-962.

50. Du P, Gu S, Jiao Y. 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. 2014;15:3495-3506.

51. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Cur Proteom. 2009;6:262-274.

52. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014;456:53-60.

53. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015;11:2620-2634.

54. Chen W, Tang H, Ye J, Lin H, Chou KC. iRNA-PseU: Identifying RNA pseudouridine sites Molecular Therapy. Nucl Aci. 2016;5:332.

55. Liu B, Fang L, Long R, Lan X, Chou KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinform. 2016;32:362-369.

56. Liu B, Long R, Chou KC. iDHS-EL: Identifying DNase I hypersensitivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinform. 2016;32:2411-2418.

57. Feng P, Ding H, Yang H, Chen W, Lin H, Chou KC. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Molecular Therapy. Nucl Aci. 2017;7:155-163.

58. Liu B, Wang S, Long R, Chou KC. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinform. 2017;33:35-41.

59. Liu B, Yang F, Chou KC. 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Molecular Therapy. Nucl Aci. 2017;7:267-277.

60. Sabooh MF, Iqbal N, Khan M, Khan M, MaqboolHF. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J Theor Biol. 2018;452:1-9.

61. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucl Aci Res. 2015;43:65-71.

62. Liu B, Wu H, Chou KC. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci. 2017;9:67-91.

63. Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinform. 2018;34:4196-4204.

64. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Molecular Therapy. Nucl Aci. 2018;11:468-474.

65. Yang H, Qiu WR, Liu G, Guo FB, Chen W, Chou KC, et al. iRSpot-Pse6NC: Identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC. Int J Biolo Sci. 2018;14:883-891.

66. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem. 2002;277:45765-45769.

67. Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J. 2003;84:3257-3263.

68. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Chapter 3, Cambridge University Press 2000.

69. Breiman L. Random forests, Machine learning. 2001;45:5-32.

70. Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, Sridharan S, et al. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol. 2011;230:56-62.

71. Pugalenthi G, Kandaswamy KK, Chou KC, Vivekanandan S, Kolatkar P, RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. Prote Pep Let. 2012;19:50-56.

72. Xu Y, Ding J, Wu LY, Chou KC. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE. 2013; 55844.

73. Jia J, Liu Z, Xiao X, Chou KC. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol. 2015;377:47-56.

74. Jia J, Liu Z, Xiao X, Liu B, Chou KC. 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. 2016;34:1946-1961.

75. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016;497:48-56.

76. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016;34558-34570.

77. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget. 2016;7:44310-44321.

78. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinforma. 2016;32:3116-3123.

79. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget. 2016;7:51270-51283.

80. Xiao X, Ye HX, Liu Z, Jia JH, Chou KC. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget. 2016;7:34180-34189.

81. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol Informa. 2017;36:1600010.

82. Liu B, Yang F, Huang DS, Chou KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinforma. 2018;34:33-40.

83. Jia J, Liu X, Qiu W, Xiao X, Chou KC. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. J Theo Biol. 2019;460:195-203.

84. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amin Aci. 2007;33:423-428.

85. Chou KC. Prediction of signal peptides using scaled window. Peptid. 2001;22:1973-1979.

86. Xu Y, Shao XJ, Wu LY, Deng NY, Chou KC. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. Peer J. 2013;1:171.

87. Xu Y, Wen X, Shao XJ, Deng NY, Chou KC. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014;15:7594-7610.

88. Qiu WR, Xiao X, Lin WZ, Chou KC. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. Biomed Res Int. 2014;947416.

89. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucle Aci Res. 2014;42:12961-12972.

90. Xu Y, Wen X, Wen LS, Wu LY, Deng NY, Chou KC. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE. 2014;9:105018.

91. Xiao X, Min JL, Lin WZ, Liu Z, Cheng X, Chou KC. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J Biomol Struct Dyn. 2015;33:2221-2233.

92. Zhang CJ, Tang H, Li WC, Lin H, Chen W, Chou KC. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget. 2016;7:69783-69793.

93. Liu Z, Xiao X, Yu DJ, Jia J, Qiu WR, Chou KC. pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. Anal Biochem. 2016;497:60-67.

94. Chen W, Ding H, Feng P, Lin H, Chou KC. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7:16895-16909.

95. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecu. 2016;21:95.

96. Qiu WR, Jiang SY, Sun BQ, Xiao X, Cheng X, Chou KC. iRNA-2methyl: identify RNA 2▯-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Medic Chem. 2017;13:734-743.

97. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 2017;8:4208-4217.

98. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep. 2017;7:42362.

99. Jia J, Zhang L, Liu Z, Xiao X, Chou KC. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinforma. 2016;32:3133-3141.

100. Ehsan A, Mahmood K, Khan YD, Khan SA, Chou KC, A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. Scien Repo. 2018;8:1039.

101. Wang J, Li J, Yang B, Xie R, Marquez-Lago TT, Leier A, et al. Bastion3: a two-layer approach for identifying type III secreted effectors using ensemble learning. Bioinform. 2018;35:2017-2028.

102. Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinforma. 2018;34:4223-4231.

103. Li F, Wang Y, Li C, Marquez-Lago TT, Leier A, Rawlings ND. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. Brief in Bioinform. 2018.

104. Song J, Wang Y, Li F, Akutsu T, Rawlings ND, et al. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief in Bioinform. 2018;20:638-658.

105. Chen W, Ding H, Zhou X, Lin H, Chou KC. iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem. 2018;59-65.

106. Khan YD, Jamil M, Hussain W, Rasool N, Khan SA, Chou KC. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. J Theor Biol. 2019;463:47-55.

107. Cheng X, Lin WZ, Xiao X, Chou KC. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinforma. 2019;398-406.

108. Chou KC. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. Curr Med Che. 2019;26:4918-4943.

109. Chou KC, Cheng X, Xiao X. pLoc_bal-mEuk: predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset. Med Chem. 2019;472-485.

110. Ehsan A, Mahmood MK, Khan YD, Barukab OM, Khan SA, Chou KC. iHyd-PseAAC (EPSV): Identify hydroxylation sites in proteins by extracting enhanced position and sequence variant feature via Chou's 5-step rule and general pseudo amino acid composition. Curr Geno. 2019;20:124-133.

111. Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC. iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomic. 2019;111:96-102.

112. Hussain W, Khan SD, Rasool N, Khan SA, Chou KC. SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Anal Biochem. 2019;568:14-23.

113. Hussain W, Khan YD, Rasool N, Khan SA, Chou KC. SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. J Theor Biol. 2019;468:1-11.

114. Lu Y, Wang S, Wang J, Zhou G, Zhang Q, Zhou X, et al. An Epidemic Avian Influenza Prediction Model Based on Google Trends. Lett Org Che. 2019;16:303-310.

115. Shen HB, Chou KC. Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J Theo Biolo. 2010;264:326-333.

116. Chou KC, Shen HB. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. PLoS ONE. 2010;5:11335.

117. Shen HB, Chou KC. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc. Anal Biochem. 2009;394:269-274.

118. Wu ZC, Xiao X, Chou KC. iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. Prot Pep Lett. 2012;19:4-14.

119. Lin WZ, Fang JA, Xiao X, ChouKC. iLoc-Animal: A multi-label

learning classifier for predicting subcellular localization of animal proteins. Mole BioSys. 2013;9:634-644.

120. Wu ZC, Xiao X, Chou KC. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mole BioSys. 2011;7:3287-3297.

121. Cheng X, Xiao X, Chou KC. pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. Mole BioSys. 2017;13:1722-1727.

122. Cheng X, Xiao X, Chou KC. pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene. 2018;644:315-321.

123. Cheng X, Zhao SG, Lin WZ, Xiao X, Chou KC. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinform. 2017;33:3524-3531.

124. Xiao X, Cheng X, Su S, Nao Q, Chou KC. pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Nat Sci. 2017;9:330-349.

125. Cheng X, Xiao X, Chou KC. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genom. 2018;110:50-58.

126. Cheng X, Xiao X, Chou KC. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genom. 2018;110:231-239.

127. Cheng X, Xiao X, Chou KC. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinform. 2018;34:1448-1456.

128. Pacharawongsakda E, Theeramunkong T. Predict Subcellular Locations of Singleplex and Multiplex Proteins by Semi-Supervised Learning and Dimension-Reducing General Mode of Chou's PseAAC. IEEE Trans Nanobiosci. 2013;12:311-320.

129. Cao JZ, Liu WQ, Gu H. Predicting Viral Protein Subcellular Localization with Chou's Pseudo Amino Acid Composition and Imbalance-Weighted Multi-Label K-Nearest Neighbor Algorithm. Prot PepLet. 2012;19:1163-1169.

130. Li LQ, Zhang Y, Zou LY, Zhou Y, Zheng XQ. Prediction of Protein Subcellular Multi-Localization Based on the General form of Chou's Pseudo Amino Acid Composition. Prot Pept Let. 2012;19:375-387.

131. Mei S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. J Theor Biol. 2012;310:80-87.

132. Huang C, Yuan J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. Biosystems. 2013;113:50-57.

133. Wang X, Li GZ, Lu WC. Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition. Prot Pep Let. 2013;20:309-317.

134. Mandal M, Mukhopadhyay A, Maulik U. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med biolo eng comp. 2015;52:331-344.

135. Cheng X, Zhao SG, Xiao X, Chou KC. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinform. 2017;33:341-346.

136. Cheng X, Zhao SG, Xiao X, Chou KC. Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 2017;8:58494-58503.

137. Xiao X, Wang P, Lin WZ, Jia JH, Chou KC. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013;436:168-177.

138. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems, Molecular Biosystem. 2013;9:1092-1100.

139. Chou KC, Zhang CT. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol. 1995;30:275-349.

140. Mohabatkar H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Prot Pep Let. 2010;17:1207-1214.

141. Zhou GP. Subcellular location prediction of apoptosis proteins. Proteins: Struct Funct Genet. 2003;50:44-48.

142. Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comp Biol Chem. 2010;34:320-327.

143. Zia-ur-Rehman, Khan A. Identifying GPCRs and their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-scale Energy Representation and Position Specific Scoring Matrix. Prot Pept Let. 2012;19:890-903.

144. Fan GL, Li QZ. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. J Theor Biol. 2013;334:45-51.

145. Huang C, Yuan JQ. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. J Theor Biol. 2013;335:205-212.

146. Hajisharifi Z, Piryaiee M, Mohammad BM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol. 2014;341:34-40.

147. Chou KC, ShenHB. Recent advances in developing web-servers for predicting protein attributes. Nat Sci. 2009;1:63-92

148. Shen HB, Chou KC. HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal Biochem. 2008;375:388-390.

149. Qiu WR, Xiao X, Chou KC. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci. 2014;15:1746-1766.

150. Liu B, Fang L, Liu F, Wang X, Chen J, Chou KC. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE. 2015;10:0121501.

151. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinform. 2017;33:2756-2758.

152. Chen Z, Zhao PY, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinform. 2018;34:2499-2502.

153. Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G, et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinform. 2018;34:684-687.

154. Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou KC, et al. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework. J Theo Biol. 2018;443:125-137.

155. Qiu WR, Sun BQ, Xiao X, Xu ZC, Jia JH, Chou KC. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genom. 2018;110:239-246.

156. Liu LM, Xu Y, Chou KC. iPGK-PseAAC: identify lysine

phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem. 2017;13:552-559.

157. Qiu WR, Jiang SY, Xu ZC, Xiao X, Chou KC. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget. 2017;8:41178-41188.

158. Wang J, Yang B, Leier A, Marquez-Lago TT, Hayashida M, Rocker A, et al. Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinform. 2018;34:2546-2555.

159. Xu Y, Li C, Chou KC. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med Chem. 2017;13:544-551.

160. Liu B, Wu H, Zhang D, Wang X, Chou KC. Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget. 2017;8:13338-13343.

161. Chou KC, Jiang SP, Liu WM, Fee CH. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinic. 1979;22:341-358.

162. Chou KC, Forsen S. Graphical rules for enzyme-catalyzed rate laws. Biochem J. 1980;187:829-835.

163. Chou KC, Forsen S, Zhou GQ. Three schematic rules for deriving apparent rate constants. Chemica Scripta. 1980;16:109-113.

164. Chou KC, Carter RE, Forsen S. A new graphical method for deriving rate equations for complicated mechanisms. Chemic Script. 1981;18:82-86.

165. Chou KC, Forsen S. Graphical rules of steady-state reaction systems. Can J Chem. 1981;59:737-755.

166. Zhou GP, Deng MH. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem J. 1984;222:169-176.

167. Chou KC. Graphic rules in steady and non-steady enzyme kinetics. J Biol Chem.1989;264:12074-12079.

168. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem. 1993;268:6119-6124.

169. Chou KC. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys Chem. 1990;35:1-24.

170. Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, Kezdy FJ, et al. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem. 1993;268:14875-14880.

171. Chou KC. Graphic rule for drug metabolism systems. Curr Drug Metabol. 2010;11:369-378.

172. Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J Theor Biol. 2011;284:142-148.

173. Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, Kezdy FJ, et al. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochem. 1993;32:6548-6554.

174. Chou KC, Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams. Nat Sci. 2011;3862-865

175. Chou KC, Forsen S. Diffusion-controlled effects in reversible enzymatic fast reaction system: Critical spherical shell and proximity rate constants. Biophys Chem. 1980;12:255-263.

176. Chou KC, Li T T, Forsen S. The critical spherical shell in enzymatic fast reaction systems. Biophys Chem. 1980;12:265-269.

177. Shen HB, Song JN, Chou KC. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng. 2009;2:136-143.

178. Chou KC, Chen NY, Forsen S. The biological functions of low-frequency phonons: 2. Cooperative effects. Chemic Script. 1981;18:126-132.

179. Chou KC. Review: Low-frequency collective motion in biomacromolecules and its biological functions. Biophy Chem. 1988;30:3-48.

180. Chou KC, Kezdy FJ, Reusser F. Review: Kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem. 1994;221:217-230.

181. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC. Using cellular automata to generate Image representation for biological sequences. Amin Aci. 2005;28:29-35.

182. Wu ZC, Xiao X, Chou KC. 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. 2010;267:29-34.