MINI REVIEW ARTICLE

A Mini-review of the Computational Methods Used in Identifying RNA 5-Methylcytosine Sites

Jianwei Li[1,2] , Yan Huang[1] and Yuan Zhou[2,*]

[1]*Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin, China;* [2]*Department of Biomedical Informatics, School of Basic Medical Sciences, Center for Noncoding RNA Medicine, Peking University, Beijing, China*

ARTICLE HISTORY

Received: October 24, 2019
Revised: January 17, 2020
Accepted: January 31, 2020

DOI:
10.2174/1389202921666200219124951

Abstract: RNA 5-methylcytosine (m^5C) is one of the pillars of post-transcriptional modification (PTCM). A growing body of evidence suggests that m^5C plays a vital role in RNA metabolism. Accurate localization of RNA m^5C sites in tissue cells is the premise and basis for the in-depth understanding of the functions of m^5C. However, the main experimental methods of detecting m^5C sites are limited to varying degrees. Establishing a computational model to predict modification sites is an excellent complement to wet experiments for identifying m^5C sites. In this review, we summarized some available m^5C predictors and discussed the characteristics of these methods.

Keywords: 5-methylcytosine, RNA modification, machine learning, prediction, sequence encoding, pseudo dinucleotide composition.

1. INTRODUCTION

RNA 5-methylcytosine (m^5C) is a pervasive regulatory mark in both eukaryotes and prokaryotes [1, 2]. Recent advances in m^5C mapping technologies have verified the presence of m^5C in tRNA, rRNA, mRNA and lncRNA, which revived the community's interest in studying the functions and mechanisms of m^5C [3, 4]. In tRNAs, m^5C is known to influence both structural and metabolic stabilization [5]. Lack of modified residues may reduce conformational stability that leads to degradation of tRNAs [6-8].

Ribosomal RNA m^5C modification is widespread in all kingdoms of life and they are usually evolutionarily conserved in genomes [1]. Recent studies suggest that m^5C methylation status is involved with the efficiency of the nuclear export of mRNA by affecting the activity of nuclear factor ALYREF/THOC4 [9, 10]. It is also reported that m^5C influences the translation of proteins. Modifications that occur at different positions may either promote or inhibit translation efficiency [11, 12]. However, the regulatory functions of m^5C are still not fully understood.

A fundamental step for further research on the functions and mechanisms of RNA m^5C modification is to obtain the position of the modification sites. Up to now, several experimental methods have been proposed to detect RNA m^5C sites, including bisulfite sequencing, meRIP-seq, Aza-IP and meCLIP [13-16]. Here we introduced these four methods briefly: 1) Bisulfite sequencing: This method is based on the different reactions between m^5C and the general cytosine when they are exposed to a sodium bisulfite environment. The general cytosine will convert into uridines with the deamination of sodium bisulfite while the m^5C will not be affected. 2) MeRIP-seq: This method was first applied to detect m^6A methylome in RNA. The antibody against m^5C will pull down the m^5C-containing RNA fragments during the immunoprecipitation and locate the m^5C at the transcription level. 3) Aza-IP: In the process of methyltransferase catalysis, the cysteine of methyltransferase forms a temporary covalent bond with the modified cytosine at carbon 6, after which the methyltransferase reverts to the free enzyme state. Aza-IP incorporates 5-azacytidine, a cytidine analog, into RNA and this compound can prevent the separation of the complex mentioned above. In this process, a C-to-G transversion occurs at methyltransferases targeted sites, which allows precision identification of m^5C modification sites. 4) miCLIP.: Unlike Aza-IP, which "trapped" methyltransferase by 5-azacytidine, miCLIP directly mutated the cysteine residues in methyltransferase into alanine, so as to achieve the goal to sequester methyltransferase after its binding to the modification sites. The reverse transcription will be terminated at the -1 position of the methylated site because of the hindrance from enzyme-RNA cross-binding.

Although the experimental methods for detecting m^5C sites have convincing accuracy, they have limitations more or less. For instance, bisulfite sequencing needs alkaline conditions that may cause RNA degradation, which is an obstacle for subsequent reverse transcription. Another defect is the low conversion rate of cytosine in RNA stem region. MeRIP-seq also suffers from detecting folded RNA secondary structures. A limitation of Aza-IP is the nonquantitative transversion of C-to-G because of the toxicity of 5-azaC. As

*Address correspondence to this author at the Department of Biomedical Informatics, School of Basic Medical Sciences, Center for Noncoding RNA Medicine, Peking University, Beijing, China; Tel/Fax: +86 10 82801585; E-mail: zhouyuanbioinfo@hsc.pku.edu.cn

for miCLIP, it relies on the mutation rate of methyltransferase but may cause changes in methylation patterns in the process of keeping a high rate of mutation. In addition, high-throughput wet experiments are all laborious and time-consuming. For the most commonly used bisulfite sequencing, it usually costs more than one month for sample preparation and sequencing, and costs more than 1000 dollars as per our collaborative company.

Based on this situation, several computational predictors for identifying RNA m^{5}C sites have been established and Fig. (**1**) shows the workflow of these predictors. Typically, three essential elements are required to constitute a sequence-based RNA m^{5}C site predictor, *i.e*., training dataset (where the known RNA m^{5}C sites settle), sequence feature encoding strategy (how to describe the proximal sequences of RNA m^{5}C sites as mathematical formulation), and the machine-learning algorithms (how to classify the sites based on the sequence features). As summarized in Table **1**, all of the available predictors were established with popular machine-learning algorithms like support vector machine (SVM) and random forest (RF). Nevertheless, they are substantially differed in the sequence feature encoding strategy and training dataset used. In this review, we provided an overview of these computational approaches from the perspective of the training dataset, encoding strategy and machine learning algorithm implementation.

2. TRAINING DATASET

Intuitively, the training dataset is composed of positive samples (known as m^{5}C site) and negative samples (non-modified sites). However, technically, the selection of training samples is not a trivial issue for building m^{5}C site predictors. On the one hand, as what is introduced above, there are several distinct high-throughput sequencing methods for m^{5}C identification, and each method has its unique principle and protocol. Besides, RNA modification like m^{5}C often exhibits species, tissue and cell-type specificity. Therefore, a set of mixed positive samples from different organisms, tissues and experiment types is preferred to assemble a robust training set. Early studies adopted human m^{5}C sites identified by bisulfite sequencing [13], which were collected from the early version of RMBase [22]. However, as demonstrated by Li *et al.*, predictors trained on this dataset often show noticeably reduced accuracy on heterogeneous datasets mixing m^{5}C site data from different experimental studies in Gene Expression Omnibus (GEO) [19]. Notably, the updated RMBase V2.0 has been released recently [23], providing a much enriched source of known m^{5}C sites. Therefore, more updated, mixed m^{5}C datasets from GEO and/or RMBase V2.0 would be a good choice to train and test the m^{5}C site predictors. On the other hand, there is still no golden standard negative samples for m^{5}C site prediction. Most predictors, if not all, use randomly picked cytosine residues that have not yet been reported to be modified as the negative samples. However, such randomly picked negative samples should include a considerable fraction of false negatives, *i.e*., the potential m^{5}C sites that have not been identified in one particular experiment. To this end, a mixed dataset incorporating more experimental m^{5}C site data could be helpful to rule out some false negatives in the training datasets. Nevertheless, a golden standard negative set identified with a novel experimental design is still highly demanded. For example, a microarray-based experimental design, which quantifies the relative m^{5}C methylation level at each site, might be useful to identify the sites that are rarely modified as the true negative sites.

To further illustrate the challenges from the heterogeneous nature of current m^{5}C site data, we here introduced a newly released m^{5}C dataset from GEO (GSE90963) as the testing set [24]. This testing dataset is released after the publication of all the abovementioned predictors, therefore it would be mostly independent of the training dataset of previous predictors. ROC (receiver operating characteristic) curve is generally used to show the performance of predictors in machine learning area and it can intuitively and accurately reflect the relationship between sensitivity and specificity, but it is not applicable here due to the lack of variable threshold of most m^{5}c site predictors. In order to directly compare the predictors' recognition ability of positive and negative samples, we took one positive and one negative sample from each transcript to verify whether these predictors can recognize both of them. We picked 333 author-reported high-confidence m^{5}C sites and corresponding 333 randomly picked non-modified sites from the new benchmark dataset. Then this independent test set was used to preliminarily evaluate the performance of the predictors and the result is shown in Table **2**. Notably, iRNA-PseColl correctly identified most of the m^{5}C sites, but it also wrongly assigned most of the non-modified sites into false positives. M5C-HPCR was more balanced than iRNA-PseColl in terms of true positive rate and false positive rate, but it still wrongly recognized more than half of the negative samples. Since we only adopted a 1:1 positive-to-negative ratio for this preliminary benchmarking test, the false positive rate will be even higher when running predictions on real RNA sequences where the number of non-modified sites overwhelms that of modified sites. It is noticeable that the size of training data for these methods is relatively small (Table **1**), indicating large and heterogeneous training dataset is important to improve the robustness of the predictor and reduce the false positive rate. By contrast, RNAm5Cfinder turns out to be just too conservative, recognizing only 36.9% of m^{5}C sites. One reason is that the RNAm5Cfinder used a training dataset with an extremely imbalanced positive-to-negative ratio (1:30), which made it robust in ruling out false positives but insensitive to true positive on the other hand. The one that showed the best overall performance in this testing set was PEA-m5C. It is a predictor for plant m^{5}C sites. As the m^{5}C sites in this new testing dataset are of high-confidence and therefore more likely to be evolutionarily conserved, it is unsurprising to observe the good performance of PEA-m5C. Nevertheless, the ability of this predictor to find evolutionarily less conserved, species-specific m^{5}C sites remains to be evaluated.

However, in the independent test dataset, there is a possibility that those negative samples may be the potential positive samples that have not yet been identified. Therefore, we then utilized two evaluation methods to estimate the proportion of potential positive samples among the negative dataset and obtained the range values of the performance statistics for a better evaluation of the predictors [25]. For the conservative estimation, the ratio of 8.5%, which was the

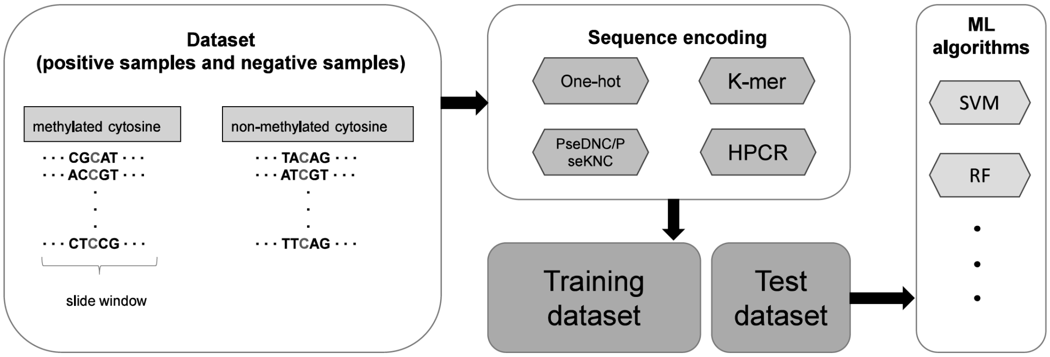

Fig. (1). Workflow of the computational pipeline to predict RNA m^{5}C sites. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 1. The dataset, encoding strategy and corresponding machine-learning algorithm of the predictors.

Name of the Predictors or Platforms	Coding Strategy	Machine Learning Algorithm	Training Dataset Source	Species	URL
iRNAm5C-PseDNC [17]	PseDNC encoding	RF	RMBase (475 positive samples)	Mostly Human and Mouse	http://www.jci-bioinfo.cn/iRNAm5C-PseDNC (server)
iRNA-PseColl [18]	PseKNC encoding	SVM	SRA027832 (120 positive samples)	Human	http://lin.uestc.edu.cn/server/iRNA-PseColl (server)
RNAm5Cfinder [19]	One-hot encoding	RF	GSE90963 GSE93749 GSE83432 (19798 positive samples)	Human Mouse	http://www.rnanut.net/rnam5cfinder (server)
PEA-m5C [20]	One-hot encoding mer encoding PseDNC encoding	RF	GSE80054 (1196 positive samples)	Arabidopsis	https://github.com/cma2015/PEA-m5C (source code)
m5C-HPCR [21]	HPCR	SVM	SRA027832 (120 positive samples)	Human	http://cslab.just.edu.cn:8080/M5C-HPCR (server)

Table 2. Observed performance of the independent test of predictors.

Predictors or Platforms	Sensitivity	Specificity	ACC	MCC
iRNAm5C-PseDNC	Website inaccessible	Website inaccessible	Website inaccessible	Website inaccessible
iRNA-PseColl	0.925	0.081	0.503	0.011
RNAm5Cfinder	0.369	0.766	0.568	0.147
PEA-m5C	0.426	0.784	0.605	0.225
M5C-HPCR	0.631	0.423	0.527	0.055

proportion of high-confidence m^{5}C sites to all candidate sites in the GSE90963 dataset, was set to represent the possibility that a negative sample is a potentially positive sample. And a radical estimation was also employed to evaluate the prediction result, where all of the negative samples that were not recognized by predictors were considered to be the potential positive samples. And the result of the estimated performance ranges is shown in Table **3**. Due to the consideration of false negatives in the dataset, the performance statistics of the predictors would show varying degrees of changes. As expected, iRNAm5C-PseDNC has the best sensitivity but unsatisfactory specificity among four predictors. This also indicates that iRNAm5C-PseDNC is more suitable when users prefer to find as many m^{5}C sites as possible, regardless of the false positive rate. On the other hand, PEA-m5C, followed by RNAm5Cfinder, showed the tightest ranges of MCC, suggesting their overall robust performance as compared with the other two predictors. Nevertheless, all predictors exhibited compromised performance on the new testing dataset. Thus, the importance of timely update of the m^{5}C site datasets is again emphasized, in order to establish robust m^{5}C predictors.

3. METHODS

3.1. Sequence Encoding Strategy

To train a machine learning model, a key procedure is to translate the RNA sequence flanking the modified/non-modified sites into a numeric form. In this section, we summarized several sequence encoding strategies, including: 1) one-hot encoding, which is a group of bits that all bits are '0' except one '1'. The '1' in different places represents different states. It is usually applied to process natural language and indicate state characteristics. 2) k-mer encoding: k-mers are the subsequences of k-length in a biological sequence. The normalized frequency of each kind of k-mer consists of the k-mer encoding. It is widely used in genome assembly, clustering and capturing nucleotide or protein sequences' features [26-28]. 3) PseDNC/PseKNC:compared with k-mer, PseDNC/PseKNC considered more global information and introduced the physical and chemical properties of nucleotides. PseDNC/PseKNC is also a popular sequence encoding strategy especially in DNA or RNA modifications. 4) HPCR: HPCR adopted the same sequence encoding strategy as PseDNC but the difference is the choice of physical-chemical properties of nucleotides. HPCR utilizes the heuristic algorithm to select the optimal set of properties among 23 properties as the parameters of PseDNC. Fig. (**2**) shows a brief illustration of each kind of encoding strategy.

3.1.1. One-hot Encoding

The one-hot encoding is a sample strategy which uses n-bit state registers to encode n states. Each state has its own register bit and only one register is valid at any time. In the process of nucleotide base sequence encoding, four bits of 0 or 1 represent four kinds of nucleotide. The A, G, C, U are translated into vectors of (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) and (0, 0, 0, 1), respectively [19, 20]. Although one-hot encoding is simple, it is easy to use and often show a substantial contribution to the prediction accuracy when the position-specific sequence propensity is prominent around the modified sites.

3.1.2. k-mer

The frequency distribution of short sequence fragments composed by nucleotide is stable in the whole genome and carries certain features of the concerned sequences. For an n-length sequence, k-mer cuts the sequence into *n*-*k*+1 substrings containing *k* nucleotides. And the feature vector could be expressed as:

$$\mathrm{F}=\left(f_1, f_2, f_3, \cdots, f_{4^k}\right) \tag{1}$$

where f_i represents the frequency of the corresponding feature.

Table 3. Estimated performance ranges of the independent test of predictors by considering potential false negatives in the dataset.

Predictors or Platforms	The Range of Sensitivity (Conservative/radical)	The Range of Specificity (Conservative/radical)	The Range of ACC (Conservative/radical)	The Range of MCC (Conservative/radical)
iRNAm5C-PseDNC	Website inaccessible	Website inaccessible	Website inaccessible	Website inaccessible
iRNA-PseColl	0.925–0.959 / 0.925–0.961	0.081–0.509 / 0.081–1.000	0.503–0.923 / 0.503–0.962	0.011–0.473 / 0.011–0.706
RNAm5Cfinder	0.369–0.480 / 0.369–0.489	0.766–0.973 / 0.766–1.000	0.568–0.674 / 0.568–0.685	0.147–0.483 / 0.147–0.518
PEA-m5C	0.426–0.521 / 0.426–0.528	0.784–0.978 / 0.784–1.000	0.605–0.704 / 0.605–0.713	0.225–0.524 / 0.225–0.552
M5C-HPCR	0.630–0.758 / 0.630–0.766	0.423–0.898 / 0.423–1.000	0.527–0.791 / 0.527–0.815	0.055–0.570 / 0.055–0.640

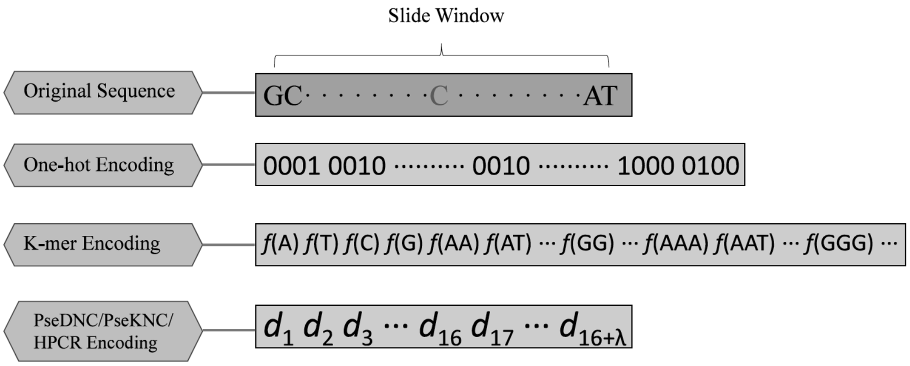
Slide Window
Original Sequence
GC· · · · · · · ·C· · · · · · · · ·AT
One-hot Encoding
0001 0010 ·········· 0010 ·········· 1000 0100
K-mer Encoding
f(A) f(T) f(C) f(G) f(AA) f(AT) ··· f(GG) ··· f(AAA) f(AAT) ··· f(GGG) ···
PseDNC/PseKNC/ HPCR Encoding
$d_1\, d_2\, d_3 \cdots d_{16}\, d_{17} \cdots d_{16+\lambda}$

Fig. (2). The overall structure of each encoding. *f* represents the frequency of the corresponding feature. *d* and *λ* are illustrated in equation (6) and (7). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

For example, when $k = 1$,

$$\mathrm{F}=\left(f(A), f(G), f(C), f(U)\right) \tag{2}$$

when $k = 2$,

$$\mathrm{F}=\left(f(AA), f(AG), f(AC), f(AU), f(GA), \cdots, f(UU)\right) \tag{3}$$

Clearly, as the value of k increase, the feature vector's dimension will rise sharply which may draw into not only an intensive computational burden but also the curse of dimensionality and/or overfitting problem. Thus, the value of k is usually not greater than 10 in practice and often concatenates feature vector obtained by using a different value of k to constitute the final input feature encoding.

3.1.3. PseDNC/PseKNC

Although k-mer encoding considers the most contiguous short sequence pattern information, it ignores the global order of such short sequences. That is to say, k-mer is less sensitive to the position-specific sequence propensity around m^{5}C modification sites. One solution could be to combine k-mer with position-specific sequence encoding like the one-hot encoding. Another is to construct a more sophisticated encoding schema to reflect local short sequence patterns and global order information at the same time. Encouraged by the successful application of pseudo amino acid composition (PseAAC) in protein/peptide sequence encoding, the pseudo dinucleotide composition (PseDNC) strategy was proposed to code RNA sequence with m^{5}C modification [17, 29]. To reflect the global sequence-order information, like PseAAC, PseDNC defined the corresponding factors as:

$$\begin{cases} \theta_1 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2}\psi(\mathrm{N}_i N_{i+1}, N_{i+1}N_{i+2}) \\ \theta_2 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3}\psi(N_i N_{i+1}, N_{i+2}N_{i+3}) \\ \theta_3 = \dfrac{1}{L-4}\displaystyle\sum_{i=1}^{L-4}\psi(N_i N_{i+1}, N_{i+3}N_{i+4}) \qquad\qquad (\lambda < L) \\ \qquad\qquad \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L-1-\lambda}\displaystyle\sum_{i=1}^{L-1-\lambda}\psi(N_i N_{i+1}, N_{i+\lambda}N_{i+\lambda+1}) \end{cases} \tag{4}$$

where N represents the nucleotide; L is the length of the sequence; θ_1 is called the first-tier correlation which reflects the sequence-order correlation between any two dinucleotides which are next to each other; θ_2 is called the second-tier correlation and it represents the sequence-order correlation between any two dinucleotides which separated by one dinucleotide; θ_3 is called the third-tier correlation and it represents the sequence-order correlation between any two dinucleotides which separated by two dinucleotides, and so on. λ represents the highest rank of the interval of two dinucleotides. Ψ is a function that reflects the correlation between two dinucleotides which can be calculated by the following equation:

$$\psi(N_i N_{i+1}, N_{i+j}N_{i+j+1}) = \\ \frac{1}{u}\sum_{k=1}^{u}\left[PC^k(N_i N_{i+1}) - PC^k(N_{i+j}N_{i+j+1})\right]^2 \tag{5}$$

where PC is the normalized value of RNA physical-chemical property for the dinucleotide composed by two adjacent nucleotides; u is the number of PC considered. As for the detection of RNA m^{5}C, three physical-chemical properties have been adopted, *i.e.*, enthalpy, entropy and free energy [30]. Finally, the mathematical expression of the sequence can be described as:

$$\mathrm{D}=\left[d_1 d_2 \cdots d_{16} d_{16+1} \cdots d_{16+\lambda}\right]^T \tag{6}$$

And the d values in the above equation can be calculated by:

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}\,(1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}\,(17 \le k \le 16+\lambda) \end{cases} \tag{7}$$

Where k is the index of d in Eq. (6); f_k is the same as f in Eq. (1); w is the weight factor range from 0 to 1. The value of λ and w should be obtained by grid search (which usually based on cross-validation performance on the training dataset) [30].

Pseudo k-tuple nucleotide composition (PseKNC) is a more general coding strategy which takes k nucleotides (rather than two nucleotides) as the input to construct the pseudo nucleotide encoding [18, 31]. Compared with one-hot or k-mer encoding, both PseDNC and PseKNC associate short sequence order information with RNA physical-chemical properties. And they are widely used in other functional site prediction tasks like predicting recombination spots, RNA splicing sites and RNA m^{6}A modification sites [18, 32-34] Besides, practically, combining multiple sequence encoding could often be helpful to predictor performance improvement. As recently demonstrated by Song *et al.* in the m^{5}C site prediction task, combining PseDNC with one-hot and k-mer encoding could further enhance the prediction accuracy [20]. Finally, PseKNC could be calculated with custom parameters at http://lin-group.cn/pseknc/default.aspx.

3.1.4. HPCR

Considering that there are at least 23 physical-chemical properties of nucleotides which are: (1) Rise [35]; (2) Roll [35]; (3) Shift [35]; (4) Slide [35]; (5) Tilt [35]; (6) Twist [35]; (7) Stacking energy [35]; (8) Enthalpy [36]; (9) Enthalpy2 [36]; (10) Entropy [36]; (11) Entropy2 [37]; (12) Free energy [37]; (13) Free energy2 [37]; (14) Adenine content [38]; (15) Cytosine content [38]; (16) GC content [38]; (17) Guanine content [38]; (18) Keto (GT) content [38]; (19) Purine (AG) content [38]; (20) Thymine content [38]; (21) Hydrophilicity [39]; (22) Hydrophilicity [39]; (23) Base tacking energy [39]. And there is no evidence proving that the enthalpy, entropy and free energy used by PseDNC are always superior to the other 20 ones. Zhang *et al.* presented a heuristic reduction algorithm to screen non-redundant, informative features to make a further improvement in the accuracy of prediction tools [21]. This method first quantifies the redundancy of each property in the context of all properties with the following equation:

$$V_{redundancy}(S, PC^{i}) = Acc(S - PC^{i}) - Acc(S) \tag{8}$$

where S is the set of the total properties; PC_i is the i-th property in S; $Acc(S)$ is the prediction accuracy considering the properties in S; If $V_{redundancy}$>=0 means that PC_i is a redundant property in set S. Then the top K properties with the highest redundancy are selected as the initial element of the K redundant subsets. Each redundant subset is expanded with an iterative procedure which selects the property with max redundancy to the corresponding rest non-redundant subset and appends it to the redundant subset, until the non-redundant subset does not contain any redundant properties (Fig. **3**). HPCR provides an efficient scheme for the rational selection of RNA physical-chemical properties and further improves the performance of the RNA m^{5}C predictor. However, since the prediction accuracy is used in assessing the feature importance and redundancy, an independent validation/testing dataset is necessary to avoid over-fitting problem induced during feature selection [40].

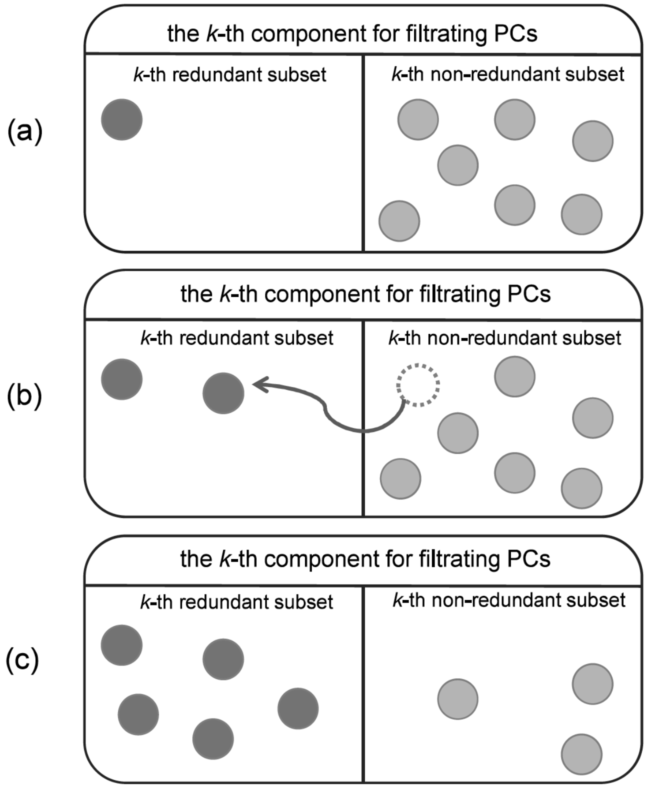
(a)
the k-th component for filtrating PCs
k-th redundant subset
k-th non-redundant subset
(b)
the k-th component for filtrating PCs
k-th redundant subset
k-th non-redundant subset
(c)
the k-th component for filtrating PCs
k-th redundant subset
k-th non-redundant subset

Fig. (3). The expansion of the *k*-th redundancy subset in HPCR encoding. (**a**) The *k*-th redundant subset only contains the initial PC, which belongs to the top *K* redundancy PC. The *k*-th non-redundant subset consists of all PCs except the one in the left subset. (**b**) The PC with the highest value of redundant in the set of the candidate will be moved to the set of redundancy. (**c**) Repeat the previous step until the set of candidate PCs does not contain any redundancy PCs, and finally, the rest members (dots on the right) of the set of candidate PCs are deeded as optimal PCs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.2. Machine-learning Algorithm

Selecting an appropriate machine-learning algorithm is another important step for the successful prediction of RNA m^{5}C sites. SVM (support vector machine) is a popular machine learning algorithm skilled in binary classification problems (*i.e.*, the problems to classify samples into two classes). It maps the input vector to a high-dimensional feature space through a non-linear function, and constructs an optimal classification hyperplane in this space, thereby maximizing the separation boundary between positive and negative samples. Most of the predictors discussed here adopt SVM as their operation engine. Especially, in view of K groups of non-redundant feature sets obtained by HPCR, M5C-HPCR adopted an ensemble learning which utilizes K SVM classifiers to fit the training data, respectively and ensembles these classifiers by a simple averaging scheme [21]. Another popular algorithm is the RF (random forest) which is adopted by RNAm5Cfinder and PEA-m5C. Random forest is an algorithm that integrates multiple decision trees through ensemble learning. It integrates all the classification voting results and specifies the category with the most votes as the final output. Benefit from the ensemble learning, the accuracy is higher than a single decision tree. Moreover, due to the randomness of samples and characteristics, random forest avoids overfitting problems to a certain extent and has certain anti-noise abilities.

As usual, k-fold cross-validation, jack-knife cross-validation and independent test can be used to examine the performance of these classifiers. *ACC* (accuracy) is the simplest and most common index that reflects the performance of a predictor. Besides, most predictors also adopt *Sn* (sensitivity), *Sp* (specificity), *Pr* (precision) and *MCC* (Matthews correlation coefficient) as their evaluation indexes. These indexes can be calculated as follows:

$$\begin{cases} Acc = \dfrac{TP+TN}{TP+FP+TN+FN} \\ Sn = \dfrac{TP}{TP+FN} \\ Sp = \dfrac{TN}{TN+FP} \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \end{cases} \tag{9}$$

where *TP*, *TN*, *FP* and *FN* represent the number of true positive, true negative, false positive and false negative samples, respectively. For a positive-to-negative balanced testing dataset, all of these evaluation indexes provide a reasonable assessment of predictor performance. However, due to the low absolute fraction of m^{5}C compared to all cytosine residues in RNA sequences, there should be much more negative samples than positive samples in a real-world application. In such cases, some indexes like accuracy and specificity will underestimate the false positive rate of predictor, and more comprehensive indexes like *MCC* are therefore recommended.

CONCLUSION

In this review, we introduced the computational methods for predicting the location of RNA m^{5}C sites, in terms of the training dataset, sequence encoding strategy and the machine-learning algorithm they used. These computational methods provide a powerful complement for traditional sequencing methods and offer great convenience for the further study of RNA m^5C modification. Especially, mixed, heterogeneous training datasets and a combination of different feature encodings would be helpful for establishing a robust RNA m^5C site predictor. However, as demonstrated by the preliminary test of the predictors in this mini-review, the prediction performance has reached a bottleneck. A prominent problem is present predictors cannot handle the relationship between false positives and false negatives and usually appear sideways. For example, iRNA-PseColl has a higher false positive rate and RNAm5Cfinder has a higher false negative rate. Therefore, before choosing a predictor, the purpose must be clear: expect more potential modification sites or more accurate sites. In general, the performance of the predictors has room for further improvement. More experimental data are still necessary to enrich the training dataset, but some other biological features like genomic features and RNA structural features may also contribute to performance improvement. On the other hand, with the accumulation of RNA m^5C data, the main challenge is moving from the identification of m^5C sites to the functional characterization of m^5C modification. We can see some clues in RNAm5Cfinder and PEA-m5C which came up with the tissue-specific m^5C predictor aiming to provide tissue specificity information of RNA m^5C modification [19, 20]. As will be readily seen, the focus in the future will shift from demonstrating where the m^5C will be to why the m^5C will be here and the functional analysis of the RNA m^5C modification.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This study was supported by the National Natural Science Foundation of China [81672113 to J.L. and 31801099 to Y.Z.], Natural Science Foundation of Hebei Province [C2018202083 to J.L.] and the Fundamental Research Funds for the Central Universities [BMU2017YJ004 to Y.Z.].

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We appreciate the researchers who shared their m^5C site data.

REFERENCES

[1] Motorin, Y.; Lyko, F.; Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res.,* **2010**, *38*(5), 1415-1430.
http://dx.doi.org/10.1093/nar/gkp1117 PMID: 20007150

[2] Bohnsack, K.E.; Höbartner, C.; Bohnsack, M.T. Eukaryotic 5-methylcytosine (m^5C) RNA methyltransferases: mechanisms, cellular functions, and links to disease. *Genes (Basel),* **2019**, *10*(2), E102.
http://dx.doi.org/10.3390/genes10020102 PMID: 30704115

[3] Amort, T.; Sun, X.; Khokhlova-Cubberley, D.; Lusser, A. Transcriptome-wide detection of 5-methylcytosine by bisulfite sequencing. *Methods Mol. Biol.,* **2017**, *1562*, 123-142.
http://dx.doi.org/10.1007/978-1-4939-6807-7_9 PMID: 28349458

[4] Amort, T.; Lusser, A. Detection of 5-methylcytosine in specific poly(A) RNAs by bisulfite sequencing. *Methods Mol. Biol.,* **2017**, *1562*, 107-121.
http://dx.doi.org/10.1007/978-1-4939-6807-7_8 PMID: 28349457

[5] Helm, M. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res.,* **2006**, *34*(2), 721-733.
http://dx.doi.org/10.1093/nar/gkj471 PMID: 16452298

[6] Kadaba, S.; Krueger, A.; Trice, T.; Krecic, A.M.; Hinnebusch, A.G.; Anderson, J. Nuclear surveillance and degradation of hypomodified initiator tRNAMet in *S. cerevisiae. Genes Dev.,* **2004**, *18*(11), 1227-1240.
http://dx.doi.org/10.1101/gad.1183804 PMID: 15145828

[7] Alexandrov, A.; Chernyakov, I.; Gu, W.; Hiley, S.L.; Hughes, T.R.; Grayhack, E.J.; Phizicky, E.M. Rapid tRNA decay can result from lack of nonessential modifications. *Mol. Cell,* **2006**, *21*(1), 87-96.
http://dx.doi.org/10.1016/j.molcel.2005.10.036 PMID: 16387656

[8] Chernyakov, I.; Whipple, J.M.; Kotelawala, L.; Grayhack, E.J.; Phizicky, E.M. Degradation of several hypomodified mature tRNA species in *Saccharomyces cerevisiae* is mediated by Met22 and the 5′-3′ exonucleases Rat1 and Xrn1. *Genes Dev.,* **2008**, *22*(10), 1369-1380.
http://dx.doi.org/10.1101/gad.1654308 PMID: 18443146

[9] Schosserer, M.; Minois, N.; Angerer, T.B.; Amring, M.; Dellago, H.; Harreither, E.; Calle-Perez, A.; Pircher, A.; Gerstl, M.P.; Pfeifenberger, S.; Brandl, C.; Sonntagbauer, M.; Kriegner, A.; Linder, A.; Weinhäusel, A.; Mohr, T.; Steiger, M.; Mattanovich, D.; Rinnerthaler, M.; Karl, T.; Sharma, S.; Entian, K.D.; Kos, M.; Breitenbach, M.; Wilson, I.B.; Polacek, N.; Grillari-Voglauer, R.; Breitenbach-Koller, L.; Grillari, J. Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. *Nat. Commun.,* **2015**, *6*, 6158.
http://dx.doi.org/10.1038/ncomms7158 PMID: 25635753

[10] Yang, X.; Yang, Y.; Sun, B.F.; Chen, Y.S.; Xu, J.W.; Lai, W.Y.; Li, A.; Wang, X.; Bhattarai, D.P.; Xiao, W.; Sun, H.Y.; Zhu, Q.; Ma, H.L.; Adhikari, S.; Sun, M.; Hao, Y.J.; Zhang, B.; Huang, C.M.; Huang, N.; Jiang, G.B.; Zhao, Y.L.; Wang, H.L.; Sun, Y.P.; Yang, Y.G. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m^5C reader. *Cell Res.,* **2017**, *27*(5), 606-625.
http://dx.doi.org/10.1038/cr.2017.55 PMID: 28418038

[11] Tang, H.; Fan, X.; Xing, J.; Liu, Z.; Jiang, B.; Dou, Y.; Gorospe, M.; Wang, W. NSun2 delays replicative senescence by repressing p27 (KIP1) translation and elevating CDK1 translation. *Aging (Albany NY),* **2015**, *7*(12), 1143-1158.
http://dx.doi.org/10.18632/aging.100860 PMID: 26687548

[12] Li, Q.; Li, X.; Tang, H.; Jiang, B.; Dou, Y.; Gorospe, M.; Wang, W. NSUN2-Mediated m5C methylation and METTL3/METTL14-mediated m^6A methylation cooperatively enhance p21 translation. *J. Cell. Biochem.,* **2017**, *118*(9), 2587-2598.
http://dx.doi.org/10.1002/jcb.25957 PMID: 28247949

[13] Schaefer, M.; Pollex, T.; Hanna, K.; Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.,* **2009**, *37*(2), e12.
http://dx.doi.org/10.1093/nar/gkn954 PMID: 19059995

[14] Edelheit, S.; Schwartz, S.; Mumbach, M.R.; Wurtzel, O.; Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m^5C within archaeal mRNAs. *PLoS Genet.,* **2013**, *9*(6), e1003602.
http://dx.doi.org/10.1371/journal.pgen.1003602 PMID: 23825970

[15] Khoddami, V.; Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.,* **2013**, *31*(5), 458-464.
http://dx.doi.org/10.1038/nbt.2566 PMID: 23604283

[16] Hussain, S.; Sajini, A.A.; Blanco, S.; Dietmann, S.; Lombard, P.; Sugimoto, Y.; Paramor, M.; Gleeson, J.G.; Odom, D.T.; Ule, J.; Frye, M. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.,* **2013**, *4*(2), 255-261.
http://dx.doi.org/10.1016/j.celrep.2013.06.029 PMID: 23871666

[17] Qiu, W.R.; Jiang, S.Y.; Xu, Z.C.; Xiao, X.; Chou, K.C. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating

physical-chemical properties into pseudo dinucleotide composition. *Oncotarget,* **2017**, *8*(25), 41178-41188. http://dx.doi.org/10.18632/oncotarget.17104 PMID: 28476023

[18] Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K.C. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 155-163. http://dx.doi.org/10.1016/j.omtn.2017.03.006 PMID: 28624191

[19] Li, J.; Huang, Y.; Yang, X.; Zhou, Y.; Zhou, Y. RNAm5Cfinder: a web-server for predicting RNA 5-methylcytosine (m^5C) sites based on random forest. *Sci. Rep.,* **2018**, *8*(1), 17299. http://dx.doi.org/10.1038/s41598-018-35502-4 PMID: 30470762

[20] Song, J.; Zhai, J.; Bian, E.; Song, Y.; Yu, J.; Ma, C. Transcriptome-wide annotation of m^5C RNA modifications using machine learning. *Front. Plant Sci.,* **2018**, *9*, 519. http://dx.doi.org/10.3389/fpls.2018.00519 PMID: 29720995

[21] Zhang, M.; Xu, Y.; Li, L.; Liu, Z.; Yang, X.; Yu, D.J. Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. *Anal. Biochem.,* **2018**, *550*, 41-48. http://dx.doi.org/10.1016/j.ab.2018.03.027 PMID: 29649472

[22] Sun, W.J.; Li, J.H.; Liu, S.; Wu, J.; Zhou, H.; Qu, L.H.; Yang, J.H. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.,* **2016**, *44*(D1), D259-D265. http://dx.doi.org/10.1093/nar/gkv1036 PMID: 26464443

[23] Xuan, J.J.; Sun, W.J.; Lin, P.H.; Zhou, K.R.; Liu, S.; Zheng, L.L.; Qu, L.H.; Yang, J.H. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.,* **2018**, *46*(D1), D327-D334. http://dx.doi.org/10.1093/nar/gkx934 PMID: 29040692

[24] Khoddami, V.; Yerra, A.; Mosbruger, T.L.; Fleming, A.M.; Burrows, C.J.; Cairns, B.R. Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. USA,* **2019**, *116*(14), 6784-6789. http://dx.doi.org/10.1073/pnas.1817334116 PMID: 30872485

[25] Chen, W.M.; Danziger, S.A.; Chiang, J.H.; Aitchison, J.D. PhosphoChain: a novel algorithm to predict kinase and phosphatase networks from high-throughput expression data. *Bioinformatics,* **2013**, *29*(19), 2435-2444. http://dx.doi.org/10.1093/bioinformatics/btt387 PMID: 23832245

[26] Allman, E.S.; Rhodes, J.A.; Sullivant, S. Statistically consistent k-mer methods for phylogenetic tree reconstruction. *J. Comput. Biol.,* **2017**, *24*(2), 153-171. http://dx.doi.org/10.1089/cmb.2015.0216 PMID: 27387364

[27] Wen, J.; Zhang, Y.; Yau, S.S. k-mer sparse matrix model for genetic sequence and its applications in sequence comparison. *J. Theor. Biol.,* **2014**, *363*, 145-150. http://dx.doi.org/10.1016/j.jtbi.2014.08.028 PMID: 25158165

[28] Carvalho, A.B.; Dupim, E.G.; Goldstein, G. Improved assembly of noisy long reads by k-mer validation. *Genome Res.,* **2016**, *26*(12), 1710-1720. http://dx.doi.org/10.1101/gr.209247.116 PMID: 27831497

[29] Shen, H.B.; Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.,* **2008**, *373*(2), 386-388. http://dx.doi.org/10.1016/j.ab.2007.10.012 PMID: 17976365

[30] Feng, P.; Ding, H.; Chen, W.; Lin, H. Identifying RNA 5-methylcytosine sites *via* pseudo nucleotide compositions. *Mol. Biosyst.,* **2016**, *12*(11), 3307-3311. http://dx.doi.org/10.1039/C6MB00471G PMID: 27531244

[31] Sabooh, M.F.; Iqbal, N.; Khan, M.; Khan, M.; Maqbool, H.F. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. *J. Theor. Biol.,* **2018**, *452*, 1-9. http://dx.doi.org/10.1016/j.jtbi.2018.04.037 PMID: 29727634

[32] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **2019**, *111*(1), 96-102. http://dx.doi.org/10.1016/j.ygeno.2018.01.005 PMID: 29360500

[33] Yang, H.; Qiu, W.R.; Liu, G.; Guo, F.B.; Chen, W.; Chou, K.C.; Lin, H. iRSpot-Pse6NC: Identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.,* **2018**, *14*(8), 883-891. http://dx.doi.org/10.7150/ijbs.24616 PMID: 29989083

[34] Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.,* **2014**, *2014,* 623149. http://dx.doi.org/10.1155/2014/623149 PMID: 24967386

[35] Pérez, A.; Noy, A.; Lankas, F.; Luque, F.J.; Orozco, M. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.,* **2004**, *32*(20), 6144-6151. http://dx.doi.org/10.1093/nar/gkh954 PMID: 15562006

[36] Goñi, J.R.; Pérez, A.; Torrents, D.; Orozco, M. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.,* **2007**, *8*(12), R263. http://dx.doi.org/10.1186/gb-2007-8-12-r263 PMID: 18072969

[37] Freier, S.M.; Kierzek, R.; Jaeger, J.A.; Sugimoto, N.; Caruthers, M.H.; Neilson, T.; Turner, D.H. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA,* **1986**, *83*(24), 9373-9377. http://dx.doi.org/10.1073/pnas.83.24.9373 PMID: 2432595

[38] Friedel, M.; Nikolajewa, S.; Sühnel, J.; Wilhelm, T. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.,* **2009**, *37*(Database issue), D37-D40. http://dx.doi.org/10.1093/nar/gkn597 PMID: 18805906

[39] Barzilay, I.; Sussman, J.L.; Lapidot, Y. Further studies on the chromatographic behaviour of dinucleoside monophosphates. *J. Chromatogr. A,* **1973**, *79*, 139-146. http://dx.doi.org/10.1016/S0021-9673(01)85282-1 PMID: 4350764

[40] Ponnuswamy, P.K.; Gromiha, M.M. On the conformational stability of oligonucleotide duplexes and tRNA molecules. *J. Theor. Biol.,* **1994**, *169*(4), 419-432. http://dx.doi.org/10.1006/jtbi.1994.1163 PMID: 7526075