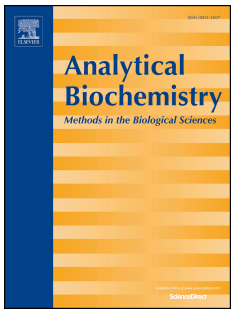
Accepted Manuscript



Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble

Ming Zhang, Yan Xu, Lei Li, Zi Liu, Xibei Yang, Dong-Jun Yu

PII: S0003-2697(18)30337-3

DOI: 10.1016/j.ab.2018.03.027

Reference: YABIO 12980

To appear in: *Analytical Biochemistry*

Received Date: 25 December 2017

Revised Date: 27 March 2018

Accepted Date: 28 March 2018

Please cite this article as: M. Zhang, Y. Xu, L. Li, Z. Liu, X. Yang, D.-J. Yu, Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble, *Analytical Biochemistry* (2018), doi: 10.1016/j.ab.2018.03.027.

[Clear Version]

Accurate RNA 5-methylcytosine Site Prediction Based on Heuristic Physical-Chemical Properties Reduction and Classifier Ensemble

Ming Zhang [a, b,*], Yan Xu [a], Lei Li [a], Zi Liu [b], Xibei Yang [a], and Dong-Jun Yu [b,*]

[a] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, China, 212003,

[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, 210094

[*] Address correspondence to M. Zhang at zhangming@just.edu.cn or D. J. Yu at njyudj@njust.edu.cn

Tel.: +86-025-84316190
Fax: +86-025-84315960

ABSTRACT

RNA 5-methylcytosine (m^5C) is an important post-transcriptional modification that plays an indispensable role in biological processes. The accurate identification of m^5C sites from primary RNA sequences is especially useful for deeply understanding the mechanisms and functions of m^5C. Due to the difficulty and expensive costs of identifying m^5C sites with wet-lab techniques, developing fast and accurate machine-learning-based prediction methods is urgently needed. In this study, we proposed a new m^5C site predictor, called M5C-HPCR, by introducing a novel heuristic nucleotide physicochemical property reduction (HPCR) algorithm and classifier ensemble. HPCR extracts multiple reducts of physical-chemical properties for encoding discriminative features, while the classifier ensemble is applied to integrate multiple base predictors, each of which is trained based on a separate reduct of the physical-chemical properties obtained from HPCR. Rigorous jackknife tests on two benchmark datasets demonstrate that M5C-HPCR outperforms state-of-the-art m^5C site predictors, with the highest values of *MCC* (0.859) and *AUC* (0.962). We also implemented the webserver of M5C-HPCR, which is freely available at http://cslab.just.edu.cn:8080/M5C-HPCR/.

Keywords: RNA 5-methylcytosine, pseudo dinucleotide composition, heuristic properties reduction, classifier ensemble

1. Introduction

Among the more than 100 post-transcriptional RNA modifications, RNA 5-methylcytosine (m^5C), catalyzed on the fifth-position carbon atom of cytosine by RNA methyltransferase (Fig. 1), is one of the most popular modifications and plays an indispensable role in biological processes [1-6]. Recent studies have demonstrated that m^5C modifications play a crucial role in affecting aminoacylation and codon identification, in stabilizing the tRNA secondary structure, in regulating stress responses, and in the proliferation of stem cells [5-10]. Therefore, accurately obtaining knowledge on m^5C sites is vitally important for both basic biomedical research and practical drug development.

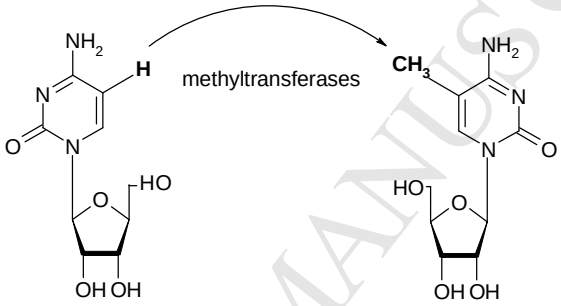


Fig. 1. Illustration of the RNA 5-methylcytosine (m^5C) modifications. The formation of m^5C is catalyzed by m^5C methyltransferases.

Much effort has been made to identify m^5C sites from RNA sequences to understand the mechanisms and functions of this post-transcriptional modification. Traditional wet-lab experimental methods, such as bisulfite-seq [5, 11], m5C-RIP [4], Aza-IP [12] and miCLIP [13], have been widely used to identify m^5C sites. However, such wet-lab methods are both expensive and time-consuming. With advanced sequencing technology and concerted genome projects, large volumes of RNA sequences have been accumulated; thus, developing intelligent computational methods for fast and accurate detection of m^5C sites from RNA sequences would be especially useful and is urgently needed.

In the past few years, a number of computational methods have been developed to predict post-translational modification sites, including 5-methylcytosine from primary RNA sequences [3, 14-25]. For example, Chen et al. developed the first m^6A site predictor, called iRNA-Methyl [16], using pseudo dinucleotide composition (PseDNC) features and a support vector machine (SVM) classifier. Liu et al. released pRNAm-PC, which is also an m^6A site predictor, by encoding RNA segments via a series of auto-covariance and cross-covariance transformations on a

physical-chemical matrix [23]. Li et al. constructed the TargetM6A, which performs m^6A sites with position-specific nucleotide propensities and an SVM [24]. Li et al. developed the first pseudouridine site predictor, termed PPUS, using binary features encoding primary RNA sequences [26]. Chen et al. further improved the prediction performance of pseudouridine sites by incorporating chemical properties and nucleotide density into the features [20]. Predictors for the m^1A site prediction also appeared recently [17].

Recently, Feng et al. [3] developed the first predictor, denoted M5C-PseDNC, which is specifically designed to identify m^5C sites from RNA sequence. M5C-PseDNC performs prediction of the m^5C sites from RNA sequences using pseudo dinucleotide composition (PseDNC) feature presentation and an SVM classifier. Qiu et al. constructed iRNAm5C-PseDNC, which is the most recently released m^5C site predictor. The iRNAm5C-PseDNC performs m^5C site prediction by incorporating additional physical-chemical properties into pseudo-dinucleotide composition [27].

Both M5C-PseDNC and iRNAm5C-PseDNC achieved promising prediction performance on m^5C sites from RNA sequences. Nevertheless, the following aspects motivate the study described in this paper.

First, M5C-PseDNC utilizes three physical-chemical properties of nucleotides, i.e., enthalpy, entropy and free energy, to encode each RNA segment into a PseDNC feature vector. However, there are many other physical-chemical properties of nucleotides that can be used, and the author did not explain the reason for choosing the three physical-chemical properties. Hence, finding the optimal physical-chemical properties for encoding more discriminative PseDNC features could help to further improve the performance of M5C-PseDNC. Second, although iRNAm5C-PseDNC reports promising performance and provides a web-server, it is found that there exists severe redundancy in the training dataset, which will lead to a deteriorating generalization capability of the trained prediction model.

The abovementioned issues motivate us to develop a more effective and applicable m^5C site predictor by optimizing the subset of physical-chemical properties for encoding features.

In our recent work [25], a heuristic nucleotide physical-chemical properties selection (HPCS) algorithm was proposed to optimize the subset of nucleotide physical-chemical properties for encoding the PseDNC feature, which has been demonstrated to be especially useful for m^6A site prediction. We also have tested HPCS on the m^5C site prediction problem, and promising results were achieved. However, on the one hand, HPCS can only obtain a single optimized physical-chemical property subset for encoding PseDNC features and does not consider the

complementarity between multiple optimized physical-chemical property subsets. On the other hand, in HPCS, the two parameters, i.e., the maximum correlation tie λ and the weight factor w, are difficult to determine [25].

In this study, we aim to further improve the prediction accuracy of m^5C sites by proposing a novel heuristic nucleotide physical-chemical property reduction (HPCR) algorithm, which can be considered an improved version of HPCS. HPCR can obtain multiple subsets, called reducts, from the original set of nucleotide physical-chemical properties by heuristically removing the property under the guidance of the improved prediction; then, on each of the obtained reducts, we train an SVM prediction model. Finally, the trained SVM models are assembled to perform the final prediction. Based on the proposed method, we implemented a new m^5C site predictor, called M5C-HPCR. We examined M5C-HPCR with rigorous jackknife tests and compared it with existing popular m^5C site predictors. Comparison results demonstrated that the proposed M5C-HPCR achieves promising m^5C site prediction with the accuracy outperforming the state-of-the-art m^5C site predictors.

2. Materials and methods

2.1 Benchmark dataset

The dataset Met1320, which was constructed by Feng et al. [3], is used in this study to demonstrate the efficacy of the proposed method. Samples in Met1320 are 41-nt RNA segments, which were detected from *H. sapiens* and have been manually checked and deposited into RMBase [6]. The Met1320 consists of a positive sample subset, denoted as S^+, and 10 negative sample subsets, denoted as $S_0^-, S_1^-, \cdots$, and S_9^-. Each positive sample is a 41-nt RNA segment with an m^5C site in the center, while each negative sample is a 41-nt RNA segment with an unmethylated cytosine in the center. Each positive or negative subset consists of 120 RNA segments and the sequence similarity between any two segments in the subset is less than 70%. More details for the construction of Met1320, refer to [3].

As mentioned in the introduction section, iRNAm5C-PseDNC [27] is the most recently released m^5C site predictor. To fairly compare the proposed method with iRNAm5C-PseDNC, the same dataset (termed as Met1900) that was used to evaluate iRNAm5C-PseDNC has also been considered in this study. Met1900 consists of 475 positive and 1425 negative samples, which are extracted from RMBase [6]. Each sample is also a 41-nt RNA segment. However, the pairwise sequence similarity between the sample sequences is not culled to 70%, which leads to the in

Met1900. In light of this, Met1900 is only used for comparison between iRNAm5C-PseDNC and the proposed M5C-HPCR method, while Met1320 is used for training the online prediction model of M5C-HPCR.

2.2 Pseudo dinucleotide composition (PseDNC)

Each RNA sample (i.e., RNA segment), denoted as R, can be formulated as follows:

$$R = N_1 N_2 \cdots N_i \cdots N_{L\text{-}1} N_L \tag{1}$$

where N_i $(1 \le i \le L)$ represents the *i*-th nucleotide in R, and L is the length of R. Clearly, each N_i belongs to one of the 4 native nucleotides, i.e.,

$$N_i \in \{\text{A (adenine), C (cytosine), G (guanine), U (uracil)}\} \tag{2}$$

How to transform each RNA sample as formulated in Eq. (1) to a feature vector with fixed length is a critical step for designing a machine-learning-based m^5C site predictor. Recently, pseudo dinucleotide composition (PseDNC) [28], which encodes both the local and global sequence pattern information of an RNA sample sequence, was introduced and has been successfully applied to the fields of computational genetics and genomics [3, 16, 29-35]. Here, we briefly introduce the PseDNC feature presentation method.

For a given RNA sample R with L nucleotides as defined in Eq. (1), the PseDNC feature vector of R can be formulated as follows:

$$\mathbf{f}_{\text{PseDNC}} = [d_1, d_2, \cdots, d_{16}, d_{16+1}, d_{16+2}, \cdots, d_{16+\lambda}]^T \tag{3}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda} \theta_i} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda} \theta_i} & (16 < k \le 16+\lambda) \end{cases} \tag{4}$$

where f_k $(1 \le k \le 16)$ is the normalized occurrence frequency of the *k*-th non-overlapping dinucleotide in the RNA sequence. The parameter λ is an integer representing the highest count tier of the correlation along the RNA sequence; w is the weight factor ranging from 0 to 1 for balancing the significance of 2-tuple nucleotide compositions and correlation factors; θ_j $(1 \le j \le \lambda)$ is the j-tier correlation factor, which reflects the sequence order correlation between all

the most contiguous dinucleotides along the RNA sequence sample, defined as follows:

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Psi(N_iN_{i+1}, N_{i+j}N_{i+j+1}) \quad (j=1,2,\cdots,\lambda;\lambda<L-1) \tag{5}$$

where the correlation function $\Psi(\cdot,\cdot)$ is given by

$$\Psi(N_iN_{i+1}, N_{i+j}N_{i+j+1}) = \frac{1}{u}\sum_{k=1}^{u}[\mathrm{PC}^k(N_iN_{i+1})\text{-}\mathrm{PC}^k(N_{i+j}N_{i+j+1})]^2 \tag{6}$$

where the *u* is the number of physical-chemical properties considered; and $\mathrm{PC}^k(N_iN_{i+1})$ represents the normalized value of the *k*-th physical-chemical property for the dinucleotide N_iN_{i+1} at the position *i* in the RNA sequence.

As far as we know, there are at least 23 kinds of physical-chemical properties that can be used to encode an RNA sequence [36-41]. Table S1 in the Supplementary Material presents the original values of the 23 physical-chemical properties. To facilitate the computation of Eq. (6), all the original values of the 23 physical-chemical properties on the 16 native dinucleotides were subjected to a standard conversion, as described by the following equation:

$$\mathrm{PC}^i(j) = \frac{\mathrm{PC}^i_{original}(j) - \mathrm{Mean}(i)}{\mathrm{Std}(i)} \tag{7}$$

where $\mathrm{PC}^i_{original}(j)$ and $\mathrm{PC}^i(j)$ ($1 \le i \le 23$, $1 \le j \le 16$) are the original and normalized physical-chemical property values, respectively, of the *i*-th physical-chemical property on the *j*-th dinucleotide type. $\mathrm{Mean}(i)$ is the mean of the original values of 16 dinucleotides for the *i*-th physical-chemical property, and $\mathrm{Std}(i)$ is the corresponding standard deviation. Table S2 in the Supplementary Material provides normalized values of the 23 physical-chemical properties. For each type of 23 physical-chemical properties, the normalized values on the 16 dinucleotides have a mean and unit variance of zero.

From Eq. (4) ~ (6), we can see that the PseDNC feature representation relies not only on parameter setting (i.e., maximum correlation tie λ and weight factor w) but also on the subset of physical-chemical properties considered. Usually, parameters of the maximum correlation tie λ and weight factor w setting can be optimized by grid search technology. However, there are 23 kinds of physical-chemical properties that can be used to encode an RNA sample sequence. How to select the subset(s) of physical-chemical properties is a problem that should be deeply studied, which will be further discussed in the subsequent section.

2.3 Physical-chemical properties reduction

In M5C-PseDNC, three physical-chemical properties of nucleotides (i.e., enthalpy, entropy and free energy) were used to extract PseDNC features for m^5C site prediction [3]. However, there is no evidence that these 3 physical-chemical properties are superior to the other 20 properties for PseDNC feature representation. In light of this, we will investigate how to extract optimized subsets of physical-chemical properties for m^5C site prediction through a heuristic reduction algorithm.

2.3.1 Measure the quality of PseDNC feature presentation

Let $S_{all} = \{\mathrm{PC}^i \mid 1 \le i \le u\}$ be a set of u known physical-chemical properties, and S be a subset of S_{all} (i.e., $S \subseteq S_{all}$). We define a *performance index*, denoted as $J_{\mathrm{PseDNC}}(S)$, to measure the prediction quality of the PseDNC feature, which is extracted based on the physical-chemical properties in S, with a prescribed prediction engine (e.g., SVM) on a given dataset.

$$J_{\mathrm{PseDNC}}(S) = \mathrm{Acc}(\mathbf{f}_{\mathrm{PseDNC}}(S)) \tag{8}$$

where $\mathbf{f}_{\mathrm{PseDNC}}(S)$ represents the PseDNC feature extracted using the physical-chemical properties in S, $\mathrm{Acc}(\mathbf{f}_{\mathrm{PseDNC}}(S))$ is the overall prediction accuracy under the feature representation of $\mathbf{f}_{\mathrm{PseDNC}}(S)$ with a prescribed prediction engine over cross-validation on a given dataset.

For each $\mathrm{PC}^i \in S$, we define a *redundancy* measure, denoted as $f_{redundancy}$, to measure the contribution of each physical-chemical property PC^i to the overall prediction accuracy as follows:

$$f_{redundancy}(S, \mathrm{PC}^i) = J_{\mathrm{PseDNC}}(S - \{\mathrm{PC}^i\}) - J_{\mathrm{PseDNC}}(S) \tag{9}$$

From the Eq. (9), we can see:

(1) If there exists a $\mathrm{PC}^i \in S$, such that $f_{redundancy}(S, \mathrm{PC}^i) > 0$, i.e., $J_{\mathrm{PseDNC}}(S - \{\mathrm{PC}^i\}) > J_{\mathrm{PseDNC}}(S)$, then we say that the PC^i is a *redundant* physical-chemical property. In this case, the quality of the PseDNC feature extracted from $S - \{\mathrm{PC}^i\}$ is better than that extracted from S, and the performance of the m^5C site predictor will be improved by removing PC^i from S.

(2) If there exists a $\mathrm{PC}^i \in S$, such that $f_{redundancy}(S, \mathrm{PC}^i) = 0$, i.e., $J_{\mathrm{PseDNC}}(S - \{\mathrm{PC}^i\}) = J_{\mathrm{PseDNC}}(S)$, then we also say that the PC^i is a *redundant* physical-chemical property because PC^i makes no contribution to the performance improvement if adding it into $S - \{\mathrm{PC}^i\}$.

(3) If there exists a $\mathrm{PC}^i \in S$, such as $f_{redundancy}(S, \mathrm{PC}^i) < 0$, i.e., $J_{\mathrm{PseDNC}}(S - \{\mathrm{PC}^i\}) < J_{\mathrm{PseDNC}}(S)$, then we say that PC^i is a *non-redundant* physical-chemical property. In this case, the quality of the PseDNC feature extracted from $S - \{\mathrm{PC}^i\}$ is worse than that extracted from S. Thus, the *non-redundant* physical-chemical property should be reserved for PseDNC feature representation.

Clearly, for a given set of physical-chemical properties S_{all}, there exists multiple distinct *reducts*, denoted as $\{S_i\}_{i=1}^{M}$, where S_i is a reduct and M is the number of reducts. We define the *optimal reduct* as S_{opt}, if $J_{\mathrm{PseDNC}}(S_{opt}) \geq \max_{1 \leq i \leq M} J_{\mathrm{PseDNC}}(S_i)$ and $S_{opt} \in \{S_i\}_{i=1}^{M}$.

Ideally, it is the best choice to obtain the *optimal reduct* of S_{all} for performing a given prediction task, e.g., m^5C site prediction in this study. However, previous studies have demonstrated that obtaining the *optimal reduct* of S_{all} is an NP-hard problem [42-45]. Hence, in the subsequent section, we will present a heuristic nucleotide physical-chemical property reduction (HPCR) algorithm by gradually removing the redundant physical-chemical properties to obtain multiple sub-optimized reducts, rather than the *optimal reduct*, of S_{all} with a lower time complexity.

2.3.2 Heuristic nucleotide physical-chemical properties reduction (HPCR) algorithm

Let $S_{all} = \{\mathrm{PC}^i \mid 1 \leq i \leq u\}$ be the set of u physical-chemical properties and $S \subseteq S_{all}$, $\mathbf{f}_{\mathrm{PseDNC}}(S)$ be the PseDNC feature extracted using physical-chemical properties in S, and K be the predefined positive integer denoting how many reducts are generated from S_{all}.

The proposed HPCR algorithm first ranks the physical-chemical properties in S_{all} according to their redundancies ($f_{redundancy}$) in descending order. Then, the top K physical-chemical properties are used to construct the K initial redundant property subsets, denoted as $\{S_{redundancy}(k)\}_{k=1}^{K}$. After that, each of the K initial redundant property subsets, i.e., $S_{redundancy}(k)$, are gradually expanded by considering the redundancies of the remaining physical-chemical properties. Finally, the K reducts, denoted as $\{S_{reduct}(k)\}_{k=1}^{K}$, can be computed using $S_{reduct}(k) = S_{all} - S_{redundancy}(k)$. We described the detailed steps of the proposed HPCR algorithm as follows:

Step 1. Rank the u physical-chemical properties:

For each physical-chemical property $\mathrm{PC}^i \in S_{all}$, we first calculate its *redundancy* using Eq. (9).

Then, the u physical-chemical properties are ranked according to their *redundancies* in descending order, denoted as $S_{Rank} = \{\mathrm{PC}_{rank}^{i}\}_{i=1}^{u}$.

Step 2. Initialize the *K* redundant physical-chemical subsets:

The top K physical-chemical properties in S_{Rank} are selected to initialize the K candidate redundant physical-chemical subsets as follows:

$$\left.\begin{array}{l} S_{redundancy}(k) \leftarrow \{\mathrm{PC}_{Rank}^{k}\} \\ S_{Rank} \leftarrow S_{Rank} - \{\mathrm{PC}_{rank}^{k}\} \end{array}\right\}, \ 1 \le k \le K \tag{10}$$

Step 3. Expand the K redundant physical-chemical subsets:

Each of the K redundant physical-chemical subsets, say $S_{redundancy}(k)$, will be gradually expanded by considering the *redundancy* measures of the remaining physical-chemical properties in $S_{all} - S_{redundancy}(k)$.

Taking the k-th initial subset, i.e., $S_{redundancy}(k)$, as an example, we expand it with an iterative procedure as follows:

For each of the remaining elements in $S_{all} - S_{redundancy}(k)$, denoted as PC^{j}, we first compute its *redundancy*, i.e., $f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^{j})$, according to Eq. (9).

Then, we can locate the j^{*}-th element, which has the maximal value of *redundancy*, from $S_{all} - S_{redundancy}(k)$ as follows:

$$j^{*} = \arg\max_{1 \le j \le N}(f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^{j})) \tag{11}$$

where $\mathrm{PC}^{j} \in S_{all} - S_{redundancy}(k)$ and $N = \left|S_{all} - S_{redundancy}(k)\right|$. $|\bullet|$ is the cardinality of a given set.

The j^{*}-th element will be added into $S_{redundancy}(k)$ (as Eq. (12)) if $f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^{j^{*}}) \ge 0$, because $\mathrm{PC}^{j^{*}}$ is still a redundant physical-chemical property in $S_{all} - S_{redundancy}(k)$:

$$S_{redundancy}(k) \leftarrow S_{redundancy}(k) \cup \{\mathrm{PC}^{j^{*}}\} \tag{12}$$

This expansion process for the k-th redundant physical-chemical property subset $S_{redundancy}(k)$ continues until $f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^{j^{*}}) < 0$ (i.e., no further redundant physical-chemical property can be found).

Step 4. Return the K reducts of S_{all}:

After the K redundant physical-chemical property subsets, i.e., $\{S_{redundancy}(k)\}_{k=1}^{K}$, have been identified, each of the K reducts of S_{all} can be commuted using Eq. (13) as follows:

$$S_{reduct}(k) = S_{all} - S_{redundancy}(k), \quad 1 \le k \le K \tag{13}$$

Algorithm 1 summarizes the detailed procedures of HPCR for extracting multiple reducts based on *redundancy* measure. Please note that the parameter K, which is a positive integer ($1 < K \le u$, u is the number of physical-chemical properties) denoting how many reducts are generated from S_{all}, needs to be prescribed before executing the algorithm.

Algorithm 1: Heuristic nucleotide physical-chemical properties reduction (HPCR) algorithm.

Input:	$S_{all} = \{\mathrm{PC}^i \mid 1 \le i \le u\}$: The set of u physical-chemical properties; K: Predefined positive integer denoting how many reducts will be generated.	
Output:	$\{S_{reduct}(k)\}_{k=1}^{K}$: A set of K reducts generated from S_{all}.	
Step 1	Rank the u physical-chemical properties in S_{all}	
	1.1	For each of the physical-chemical properties in S_{all}, calculate its redundancy using Eq. (9) as follows: $f_{redundancy}(S, \mathrm{PC}^i) = J_{\mathrm{PseDNC}}(S - \{\mathrm{PC}^i\}) - J_{\mathrm{PseDNC}}(S)$
	1.2	Rank the u physical-chemical properties according to their redundancies in a descending order: $S_{Rank} = \{\mathrm{PC}_{rank}^k\}_{k=1}^{u}$
Step 2	Initialize the K redundant physical-chemical subsets, i.e., $\{S_{redundancy}(k)\}_{k=1}^{K}$	
	2.1	FOR $k = 1, 2, \cdots, K$
	2.2	$S_{redundancy}(k) \leftarrow \{\mathrm{PC}_{Rank}^k\}$
	2.3	END FOR
Step 3	Expand the K redundant physical-chemical subsets	
	3.1	FOR $k = 1, 2, \cdots, K$
	3.2	WHILE (TRUE)
	3.3	For each of the remaining elements in $S_{all} - S_{redundancy}(k)$, denoted as PC^j, compute its *redundancy*, i.e., $f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^j)$, according to Eq. (9);
	3.4	Locate PC^{j^*}, which has the maximal value of *redundancy*, from $S_{all} - S_{redundancy}(k)$ by using Eq. (11):
	3.5	IF $f_{redundancy}(S_{all} - S_{redundancy}(k), \mathrm{PC}^{j^*}) \ge 0$
	3.6	$S_{redundancy}(k) \leftarrow S_{redundancy}(k) \cup \{\mathrm{PC}^{j^*}\}$
	3.7	ELSE
	3.8	BREAK WHILE;
	3.9	END IF
	3.10	END WHILE

	3.11	END FOR
Step 4	Return the K reducts generated from S_{all}	
	4.1	FOR $k = 1, 2, \cdots, K$
	4.2	$S_{reduct}(k) = S_{all} - S_{redundancy}(k)$
	4.3	END FOR
	4.4	RETURN $\{S_{reduct}(k)\}_{k=1}^{K}$

As for the computational efficiency, it is easy to calculate that the time complexity of Algorithm 1 is $O(K \cdot u^2)$, which is significantly better than that ($O(2^u)$) of a brute-force algorithm.

2.4 Support vector machine

Support vector machine (SVM), which was proposed by Cortes and Vapnik [46], is an efficient supervised machine-learning algorithm based on the statistical learning theory. SVM has been widely used in the realm of bioinformatics [3, 16, 17, 24, 25, 32].

The basic idea of SVM is to transform the input vectors into a high-dimension Hilbert space by kernel functions and then seek a separating hyper plane between classes with a maximal margin in this space. For more information about SVM, refer to [47-50]. In this study, the LIBSVM package (v3.20) [51, 52], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm, was taken to implement an SVM classifier. In this study, we tested different kernel functions, including linear kernel, polynomial kernel, and radial basis function (RBF) kernel, and found that RBF kernel can achieve the best performance. The regularization parameter *C* and the kernel width parameter γ of the RBF kernel were optimized based on 10-fold cross-validation using a grid search strategy in the LIBSVM package.

2.5 Classifier ensemble

Classifier ensemble, which properly integrates multiple base classifiers, has demonstrated a promising route to improve the accuracy of a prediction/classification task [53-59]. Clearly, both the ensemble strategy and the base classifier will affect the performance of an ensembled classifier. For simplicity, we used a simple averaging ensemble strategy to ensemble multiple base classifiers. As for the base classifier, SVM was adopted. The reason why SVM was taken is that we need to compare the proposed method with several existing m^5C site predictors, i.e., M5C-PseDNC [3] and M5C-HPCS [25], which also utilizes SVM as base classifier.

The K reducts generated by the HPCR algorithm may contain complementary information. In light of this, for each of the K reducts generated by HPCR, we first extracted PseDNC features

of samples based on the reduct and then trained a corresponding SVM model. Then, we ensemble the K trained SVM models using a simple averaging scheme. For the convenience of subsequent description, we term the m^5C site predictor implemented with this method as M5C-HPCR, the workflow of which is shown in Fig. 2.

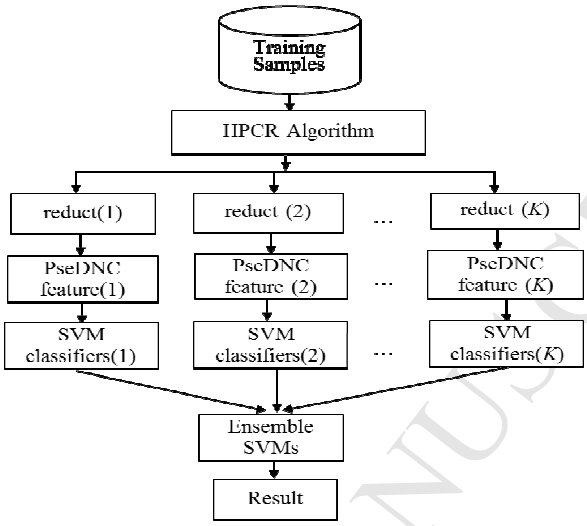


Fig. 2. Workflow of the proposed M5C-HPCR.

2.6 Performance evaluation

Four routinely used indexes in this field, i.e., specificity (*Sp*), sensitivity (*Sn*), accuracy (*Acc*), and the Matthews correlation coefficient (*MCC*) [60] were taken to evaluate the prediction performances as follows:

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

where *TP*, *FP*, *TN* and *FN* are the abbreviations of true positive, false positive, true negative and false negative, respectively.

In addition, another evaluation index *AUC*, which is the area under the receiver operating characteristic (ROC) curve and threshold independent, was exploited to evaluate the overall prediction qualities of the proposed M5C-HPCR predictor.

3. Results and discussion

3.1 Efficacy of HPCR for m^{5}C site prediction

In this section, we investigated the efficacy of the proposed HPCR (i.e., Algorithm 1) for m^{5}C site prediction. We run the HPCR algorithm by setting K=5 with SVM as the base prediction engine on the benchmark dataset obtained by merging S^+ and S_0^- in Met1320 over 10-fold cross-validation. Here, S_{all} consists of the 23 physical-chemical properties listed in Table S1 in the Supplementary Material. After running the HPCR algorithm, 5 reducts, denoted as $\{S_{reduct}(k)\}_{k=1}^{5}$, are obtained.

Table 1 summarizes the performance comparisons between the 5 reducts and S_{all} with SVM as the base prediction engine on the benchmark dataset with over 10-fold cross-validation.

Table 1. Performance comparisons between the 5 reducts and S_{all} with SVM as the base prediction engine on the benchmark dataset over 10-fold cross-validation.

Reduct	Physical-chemical properties in reduct	No. of properties	*Acc*(%)
$S_{reduct}(1)$	PC^2, PC^4, PC^5, PC^{10}, PC^{16}, PC^{17}, PC^{18}, PC^{21}, PC^{22}	9	93.33
$S_{reduct}(2)$	PC^2, PC^3, PC^4, PC^5, PC^6, PC^8, PC^{10}, PC^{12}, PC^{13}, PC^{16}, PC^{17}, PC^{18}, PC^{19}, PC^{20}, PC^{21}	15	92.92
$S_{reduct}(3)$	PC^5, PC^{13}, PC^{16}, PC^{17}, PC^{19}, PC^{21}, PC^{22}	7	92.92
$S_{reduct}(4)$	PC^2, PC^4, PC^5, PC^8, PC^{12}, PC^{14}, PC^{15}, PC^{17}, PC^{19}	9	92.92
$S_{reduct}(5)$	PC^2, PC^4, PC^5, PC^9, PC^{10}, PC^{15}, PC^{17}, PC^{19}, PC^{22}, PC^{23}	10	93.75
S_{all}	All the 23 physical-chemical properties	23	92.08
M5C-HPCR *			95.42

* M5C-HPCR denotes the model obtained by assembling the five models trained with the five reducts.

From Table 1, several observations can be made as follows:

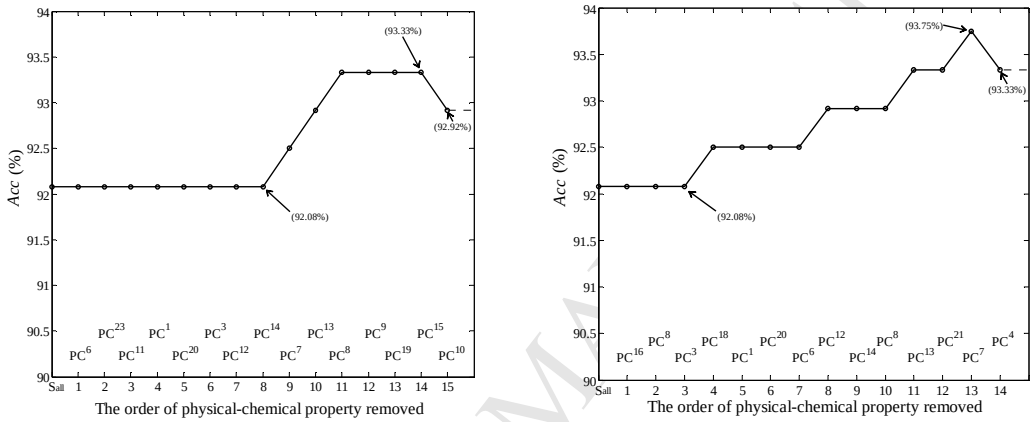
First, it is found that the five reducts constantly achieve better performance than that of S_{all}. This observation demonstrates that the proposed HPCR can remove redundant physical-chemical properties and thus help to improve the performance of m^{5}C site prediction.

Second, we found that different reducts consist of different physical-chemical properties. Hence, prediction models trained with these different reducts may contain complementary information, and the prediction performance could be further improved by assembling the models trained with different reducts. As shown in Table 1, it is found the M5C-HPCR, which is the model obtained by assembling the five models trained with the five reducts using the workflow as shown in Fig. 2, achieves the highest prediction performance with *Acc* = 95.42%.

Third, it has not escaped from our notice that several physical-chemical properties appear with high frequency among the five reducts. For example, PC^2, PC^4 and PC^{19} (i.e., roll, slide and purine (AG) content) appear in 4 out of the 5 reducts, while PC^5 and PC^{17} (i.e., tilt and guanine content) appear in all 5 reducts. This observation indicates that the physical-chemical properties appearing with high frequency are critically important for encoding discriminative PseDNC features for m^{5}C site prediction. In fact, the "Roll", "Slide" and "Tilt" properties have been demonstrated to be especially useful for nucleosome positioning prediction [32], sigma-54 promoter identification [33],

adenosine to inosine editing site prediction [14] and enhancer identification [61]. The results of these studies [14, 32, 33, 61] support our observation and demonstrate that the physical-chemical properties distilled by the proposed HPCR are useful for the identification of m^5C sites.

To deeply examine what has occurred during the reduction procedure, we randomly selected two subsets, say $S_{reduct}(1)$ and $S_{reduct}(5)$, and plotted the curve of *Acc* variation (*Acc* versus the redundant physical-chemical properties removed) during the reduction process for each of the two reducts. Fig. 3 (A) and (B) plots the curves of *Acc* variation for $S_{reduct}(1)$ and $S_{reduct}(5)$, respectively.



(A) (B)

Fig. 3. The curves of *Acc* variation (*Acc* versus the redundant physical-chemical properties removed) during the reduction process for the two obtained reducts. (A) $S_{reduct}(1)$ and (B) $S_{reduct}(5)$.

From Fig. 3, a common phenomenon is observed from the reduction process for the two reducts: the prediction performance (*Acc*) remains constant or is improved by gradually removing those redundant physical-chemical properties. This performance improvement continues until a non-redundant property appears.

Take Fig. 3 (A), the reduction process for $S_{reduct}(1)$, as an example. The *Acc* remains constant at 92.08% when the first 8 physical-chemical properties are gradually removed from S_{all}. In other words, these 8 physical-chemical properties, i.e., PC^6, PC^{23}, PC^{11}, PC^1, PC^{20}, PC^3, PC^{12} and PC^{14}, are redundant properties. The prediction performance is gradually improved from 92.08% to 93.33% and keeps constant by continually removing the subsequent 6 redundant properties, i.e., PC^7, PC^{13}, PC^8, PC^9, PC^{19} and PC^{15}. It is found the *Acc* decreases if the 15-th physical-chemical properties (i.e., PC^{10}) are removed, denoting a non-redundant property. At this point, all the redundant properties, i.e., PC^6, PC^{23}, PC^{11}, PC^1, PC^{20}, PC^3, PC^{12}, PC^7, PC^{13}, PC^8, PC^9, PC^{19} and

PC^{15} constitute $S_{redundancy}(1)$, and $S_{reduct}(1)$ can be calculated as $S_{all} - S_{redundancy}(1)$ ={ PC^{2}, PC^{4}, PC^{5}, PC^{10}, PC^{16}, PC^{17}, PC^{18}, PC^{21}, PC^{22}}={ Roll, Slide, Tilt, Entropy, GC content, Guanine content, Keto(GT) content, Hydrophilicity, Hydrophilicity2}.

Similar reduction processes for $S_{reduct}(5)$ can also be observed from Fig. 3 (B), where $S_{redundancy}(5)$ ={PC^{16}, PC^{8}, PC^{3}, PC^{18}, PC^{1}, PC^{20}, PC^{6}, PC^{12}, PC^{14}, PC^{8}, PC^{13}, PC^{21}, PC^{7}}. Accordingly, $S_{reduct}(5)$ can be calculated as $S_{reduct}(5) = S_{all} - S_{redundancy}(5)$ ={ PC^{2}, PC^{4}, PC^{5}, PC^{9}, PC^{10}, PC^{15}, PC^{17}, PC^{19}, PC^{22}, PC^{23}}={Roll, Slide, Tilt, Enthalpy2, Entropy, Cytosine content, Guanine content, Purine(AG) content, Hydrophilicity2, Base stacking energy }.

3.2 Comparisons with existing m^5C site predictors

To demonstrate the superiority of the proposed M5C-HPCR, we compared it with several state-of-the-art m^5C site predictors, including iRNAm5C-PseDNC [27], M5C-PseDNC [3] and M5C-HPCS [25], on both the Met1320 subset and Met1900 dataset.

Table 2 summarizes the comparison results between iRNAm5C-PseDNC, M5C-PseDNC, M5C-HPCS and M5C-HPCR on the Met1320 subset over the jackknife test. The results of iRNAm5C-PseDNC [27] are obtained by re-evaluating the results on the Met1320 subset over the jackknife test. The results of M5C-PseDNC are excerpted from [3]. The results of M5C-HPCS are obtained by re-evaluating HPCS algorithm [25], which was originally developed for m^6A site prediction, on the Met1320 subset over jackknife test.

Table 2 Comparison between iRNAm5C-PseDNC, M5C-PseDNC, M5C-HPCS and M5C-HPCR on Met1320 subset over jackknife test

Predictor	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	*AUC*
iRNAm5C-PseDNC [a]	81.70	95.00	88.33	0.774	0.934
M5C-PseDNC [b]	85.00	**95.83**	90.42	0.810	0.950
M5C-HPCS [c]	90.83	92.50	91.67	0.833	0.956
M5C-HPCR	**90.83**	95.00	**92.92**	**0.859**	**0.962**

[a] Results obtained by re-evaluating iRNAm5C-PseDNC [27] on the Met1320 subset over the jackknife test.

[b] Results excerpted from [3].

[c] Results obtained by re-evaluating HPCS [25] on Met1320 subset over the jackknife test.

Table 3 Comparison results between M5C-PseDNC, iRNAm5C-PseDNC, M5C-HPCR and M5C-HPCS on Met1900 dataset over jackknife test

Predictor	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	*AUC*
M5C-PseDNC [a]	84.21	94.88	92.21	0.792	0.960
iRNAm5C-PseDNC [b]	69.89	**99.86**	92.37	0.794	0.963
M5C-HPCS [c]	83.37	96.84	93.47	0.823	0.968
M5C-HPCR	**88.42**	97.33	**95.11**	**0.868**	**0.977**

[a] Results obtained by re-evaluating M5C-PseDNC [3] on the Met1320 dataset over the jackknife test.

[b] Results excerpted from [27].

[c] Results obtained by re-evaluating HPCS [25] on the Met1320 subset over the jackknife test.

Please also note that here the Met1320 subset denotes the subset obtained by merging the positive subset S^+ and the negative subset S_0^- in Met1320. The other 9 negative subsets, i.e., $S_1^- \sim S_9^-$, will be used subsequently for analyzing the robustness of the considered predictors. Table 3 compares the prediction performances between M5C-PseDNC, iRNAm5C-PseDNC, M5C-HPCR and M5C-HPCS on the Met1900 dataset over the jackknife test. Figure 4 plots the ROC curves of the four considered predictors on the Met1320 subset and the Met1900 dataset.

As shown in Table 2, we can find that the proposed M5C-HPCR achieves the highest values for 4 out of the 5 evaluation indexes, i.e., *Sn*, *Acc*, *MCC* and *AUC*. Since *MCC* and *AUC* are two indexes that measure the overall prediction quality of a model, the proposed M5C-HPCR clearly acts as the best performer with the highest values of 0.859 and 0.962 for *MCC* and *AUC*, respectively. M5C-HPCR significantly outperforms M5C-PseDNC with improvements of 4.9%, 2.5% and 1.2% in *MCC*, *Acc* and *AUC*, respectively. Compared with the second-best performer, i.e., M5C-HPCS, M5C-HPCR also performs better in *MCC*, *Acc* and *AUC* with improvements of 2.6%, 1.25% and 0.6%, respectively. These indicate that the proposed M5C-HPCR method is effective and is superior to the HPCS method, which is a state-of-the-art optimized physical-chemical properties selection algorithm.

From Table 3, the following two observations can be made:

First, we found that M5C-HPCR and M5C-HPCS again achieve better overall prediction performances (i.e., *MCC* and *AUC*) than M5C-PseDNC and iRNAm5C-PseDNC, which demonstrate the necessity to select useful physical-chemical properties for performing m^5C site prediction. In addition, M5C-HPCR outperforms M5C-HPCS with improvements of 4.5% and 0.9% for *MCC* and *AUC*, respectively, demonstrating the superiority of the proposed HPCR algorithm, which obtains multiple reducts over the HPCS algorithm, which only extracts one reduct.

Second, it was found the all four considered predictors achieve better prediction performances on the Met1900 dataset than on the Met1320 subset. The underlying reason is the higher pairwise similarity between the sequences in Met1900 as described in Section 2.1. The high pairwise similarity in a training dataset will lead to an overly optimistic evaluation of the results but will deteriorate the generalization capability of the trained prediction model. In light of this, our final

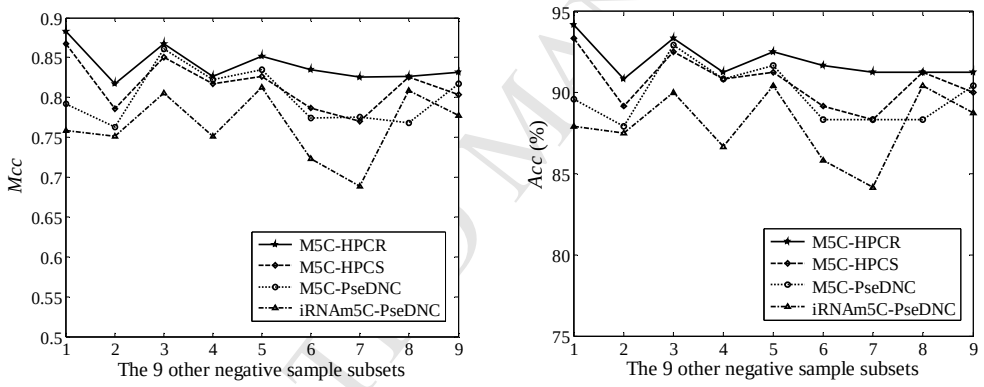
M5C-HPCR server will be trained on Met1320 rather than Met1900.

3.3 Robustness analysis of M5C-HPCR

As illustrated in Table 2, we compared the proposed M5C-HPCR with three other predictors on the subset obtained by merging the positive subset S^+ and the negative subset S_0^- in Met1320. The other 9 negative subsets, i.e., $S_1^- \sim S_9^-$, are not used for comparison.

To examine to what extent the performance of the considered predictors, including the proposed M5C-HPCR, will be affected by the selection of the negative subsets, we further compared the four considered predictors as follows: for each of the remaining 9 negative subsets, we merged it with the positive subset S^+; then, we evaluated the prediction performances of each predictor on each of the 9 merged subsets over the jackknife test.

Figure 5 plots the performance variation curves under the 9 negative subsets in Met1320 for M5C-PseDNC, M5C-HPCS, iRNAm5C-PseDNC and the proposed M5C-HPCR regarding *MCC* and *Acc*.



(A) (B)

Fig. 5. The performance variation curves under different negative subsets in Met1320 for M5C-PseDNC, iRNAm5C-PseDNC, M5C-HPCS and the proposed M5C-HPCR regarding (A) *MCC* and (B) *Acc*.

As shown in Fig. 5 (A) and (B), it was found that the proposed M5C-HPCR performs the best for 8 out of the 9 negative subsets regarding both *MCC* and *Acc*. For the eighth negative subset, M5C-HPCR achieves very comparable performances to M5C-HPCS but is still remarkably better than the other two predictors, i.e., M5C-PseDNC and iRNAm5C-PseDNC.

We also calculated the standard deviations of *MCC* and *Acc* of the four predictors over the 9 negative subsets. The standard deviations of *MCC* for M5C-HPCR, M5C-HPCS, M5C-PseDNC and iRNAm5C-PseDNC are 0.0222, 0.0315, 0.0344 and 0.0419, respectively, while that of *Acc* for the four predictors are 1.142%, 1.641%, 1.756% and 2.170%, respectively. Clearly, we can find that the proposed M5C-HPCR possesses the best stability with the smallest standard deviations of

0.0222 and 1.142% for *MCC* and *Acc*, respectively.

4. Conclusion

In this study, a new m^5C site predictor, called M5C-HPCR, is proposed based on heuristic physical-chemical properties reduction (HPCR) and classifier ensemble. The purpose of HPCR is to find multiple reducts of physical-chemical properties for encoding more discriminative features. To further improve the accuracy of m^5C site prediction, the base classifiers, each of which are trained based on a separate reduct of physical-chemical properties obtained from HPCR, are assembled.

We compared the proposed M5C-HPCR with several popular m^5C site predictors on benchmark datasets via rigorous jackknife tests. Experimental results show that the proposed M5C-HPCR outperforms state-of-the-art m^5C site predictors, including iRNAm5C-PseDNC [27], M5C-PseDNC [3] and M5C-HPCS [25]. Detailed analysis also demonstrates the robustness of the proposed M5C-HPCR. User friendly and publicly accessible web-servers represent the future direction for developing practical and useful prediction methods [62-64]. In view of this, a web-server based on HPCR algorithm has been put online available at http://cslab.just.edu.cn:8080/M5C-HPCR/. We believe that the proposed M5C-HPCR will complement existing m^5C site predictors and benefit m^5C-related research studies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61772273, 61373062, and 61572242).

References

[1] T. Amort, D. Rieder, A. Wille, D. Khokhlovacubberley, C. Riml, L. Trixl, X.Y. Jia, R. Micura, A. Lusser, Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain, Genome Biology, 18 (2017) 1-16.
[2] D. Incarnato, S. Oliviero, The RNA Epistructurome: Uncovering RNA Function by Studying Structure and Post-Transcriptional Modifications, Trends in biotechnology, 35 (2017) 318-333.
[3] P. Feng, H. Ding, W. Chen, H. Lin, Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions, Molecular Biosystems, 12 (2016) 3307--3311.
[4] S. Edelheit, S. Schwartz, M.R. Mumbach, O. Wurtzel, R. Sorek, Transcriptome-Wide Mapping of 5-methylcytidine RNA Modifications in Bacteria, Archaea, and Yeast Reveals m5C within Archaeal mRNAs, Plos Genetics, 9 (2013) 1-14.
[5] R. David, A. Burgess, B. Parker, J. Li, K. Pulsford, T. Sibbritt, T. Preiss, I.R. Searle, Transcriptome-wide Mapping of RNA 5-Methylcytosine in Arabidopsis mRNAs and non-coding RNAs, Plant Cell, (2017) doi:10.1105/tpc.1116.00751.
[6] W.J. Sun, J.H. Li, S. Liu, J. Wu, H. Zhou, L.H. Qu, J.H. Yang, RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data, Nucleic Acids Research, 44 (2016) D259-D265.
[7] P.F. Agris, Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications, Embo Reports, 9 (2008) 629-635.
[8] A. Alexandrov, I. Chernyakov, W. Gu, S. Hiley, Tr, E. Grayhack, E. Phizicky, Rapid tRNA decay can result from lack of

nonessential modifications, Molecular Cell, 21 (2006) 87-96.
[9] Y. Motorin, M. Helm, tRNA Stabilization by Modified Nucleotides, Biochemistry, 49 (2010) 4934-4944.
[10] Y. Motorin, F. Lyko, M. Helm, 5-methylcytosine in RNA: detection, enzymatic formation and biological functions, Nucleic Acids Research, 38 (2010) 1415-1430.
[11] J.E. Squires, H.R. Patel, M. Nousch, T. Sibbritt, D.T. Humphreys, B.J. Parker, C.M. Suter, T. Preiss, Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA, Nucleic Acids Research, 40 (2012) 5023-5033.
[12] V. Khoddami, B.R. Cairns, Identification of direct targets and modified bases of RNA cytosine methyltransferases, Nature Biotechnology, 31 (2013) 458-464.
[13] S. Hussain, A.A. Sajini, S. Blanco, S. Dietmann, P. Lombard, Y. Sugimoto, M. Paramor, J.G. Gleeson, D.T. Odom, J. Ule, NSun2-Mediated Cytosine-5 Methylation of Vault Noncoding RNA Determines Its Processing into Regulatory Small RNAs, Cell Reports, 4 (2013) 255-261.
[14] W. Chen, P. Feng, H. Ding, H. Lin, PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions, Scientific Reports, 6 (2016) 35123.
[15] W. Chen, P. Feng, H. Ding, H. Lin, Identifying N6-methyladenosine sites in the Arabidopsis thaliana transcriptome, Molecular Genetics & Genomics, 291 (2016) 2225-2229.
[16] W. Chen, P. Feng, H. Ding, H. Lin, K.C. Chou, iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition, Analytical Biochemistry, 490 (2015) 26-33.
[17] W. Chen, P. Feng, H. Tang, H. Ding, H. Lin, RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes, Scientific Reports, 6 (2016) 31080.
[18] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.C. Chou, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget, 8 (2016) 4208-4217.
[19] W. Chen, H. Tang, H. Lin, MethyRNA: A web-server for identification of N6-methyladenosine sites, Journal of Biomolecular Structure & Dynamics, 35 (2017) 683.
[20] W. Chen, H. Tang, J. Ye, H. Lin, K.C. Chou, iRNA-PseU: Identifying RNA pseudouridine sites, Molecular Therapy Nucleic Acids, 5 (2016) e332.
[21] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics, 33 (2017) 3518-3323.
[22] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.C. Chou, iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC, Molecular Therapy Nucleic Acids, 7 (2017) 155-163.
[23] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, K.C. Chou, pRNAm-PC: Predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties, Analytical Biochemistry, 497 (2016) 60-67.
[24] G.Q. Li, Z. Liu, H.B. Shen, D.J. Yu, TargetM6A: Identifying N6-methyladenosine Sites from RNA Sequences via Position-Specific Nucleotide Propensities and a Support Vector Machine, IEEE Trans Nanobioscience, 15 (2016) 674-682.
[25] M. Zhang, J.W. Sun, Z. Liu, M.W. Ren, H.B. Shen, D.J. Yu, Improving N 6 -methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties, Analytical Biochemistry, 508 (2016) 104-113.
[26] Y.H. Li, G. Zhang, Q. Cui, PPUS: a web server to predict PUS-specific pseudouridine sites, Bioinformatics, 31 (2015) 3362-3364.
[27] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, K.C. Chou, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget, 8 (2017) 41178-41188.
[28] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Research, 41 (2013) e68.
[29] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, K.-C. Chou, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Analytical biochemistry, 462 (2014) 76-83.
[30] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, K.-C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Analytical biochemistry, 456 (2014) 53-60.
[31] M. Kabir, M. Iqbal, S. Ahmad, M. Hayat, iTIS-PseKNC: Identification of Translation Initiation Site in human genes using pseudo k-tuple nucleotides composition, Computers in Biology & Medicine, 66 (2015) 252-257.
[32] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, K.C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics, 30 (2014) 1522-1529.
[33] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54

promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Research, 42 (2014) 12961-12972.
[34] H. Lin, Z.Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE/ACM Transactions on Computational Biology & Bioinformatics, PP (2017) 1-1.
[35] C.J. Zhang, T. Hua, W.C. Li, L. Hao, C. Wei, K.C. Chou, iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, Oncotarget, 7 (2016) 69783.
[36] A. Pérez, A. Noy, F. Lankas, F.J. Luque, M. Orozco, The relative flexibility of B-DNA and A-RNA duplexes: database analysis, Nucleic Acids Research, 32 (2004) 6144-6151.
[37] J.R. Goñi, A. Pérez, D. Torrents, Orozco, Modesto, Determining promoter location based on DNA structure first-principles calculations, Genome Biology, 8 (2007) R263.
[38] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, D.H. Turner, Improved free-energy parameters for predictions of RNA duplex stability, Proceedings of the National Academy of Sciences, 83 (1986) 9373-9377.
[39] M. Friedel, S. Nikolajewa, J. Sühnel, T. Wilhelm, DiProDB: a database for dinucleotide properties, Nucleic Acids Research, 37 (2009) D37-40.
[40] I. Barzilay, J.L. Sussman, Y. Lapidot, Further studies on the chromatographic behaviour of dinucleoside monophosphates, Journal of chromatography A, 79 (1973) 139-146.
[41] P.K. Ponnuswamy, M.M. Gromiha, On the conformational stability of oligonucleotide duplexes and tRNA molecules, Journal of Theoretical Biology, 169 (1994) 419-432.
[42] Z. Meng, Z. Shi, Extended rough set-based attribute reduction in inconsistent incomplete decision systems, Information Sciences, 204 (2012) 44-69.
[43] Y. Qian, J. Liang, W. Pedrycz, C. Dang, Positive approximation: An accelerator for attribute reduction in rough set theory, Artificial Intelligence, 174 (2010) 597-618.
[44] S. Zhao, H. Chen, C. Li, M. Zhai, RFRR: Robust Fuzzy Rough Reduction, IEEE Transactions on Fuzzy Systems, 21 (2013) 825-841.
[45] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters, 24 (2003) 833-849.
[46] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning, 20 (1995) 273-297.
[47] V.N. Vapnik, An overview of statistical learning theory, IEEE Transactions on Neural Networks, 10 (1999) 988-999.
[48] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, Journal of Biological Chemistry, 277 (2002) 45765-45769.
[49] N. Cristianini, J. Shawe-Taylor, An introduction to support Vector Machines: and other kernel-based learning methods (Printed in the United Kingdom at the University Press, 2000).
[50] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophysical Journal, 84 (2003) 3257-3263.
[51] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, Acm Transactions on Intelligent Systems & Technology, 2 (2011) 27.
[52] R.E. Fan, P.H. Chen, C.J. Lin, T. Joachims, Working Set Selection Using Second Order Information for Training Support Vector Machines, Journal of Machine Learning Research, 6 (2005) 1889-1918.
[53] C. Wei, P. Xing, Z. Quan, Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines, Scientific Reports, 7 (2017) 40242.
[54] S. Wan, Y. Duan, Q. Zou, HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source, Proteomics, 17 (2017).
[55] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, Z. Hong, Improving tRNAscan-SE Annotation Results via Ensemble Classifiers, Molecular Informatics, 34 (2015) 761.
[56] C. Lin, Y. Zou, J. Qin, X. Liu, Y. Jiang, C. Ke, Q. Zou, Hierarchical classification of protein folds using a novel ensemble classifier, Plos One, 8 (2013) e56499.
[57] H.B. Shen, K.C. Chou, Ensemble classifier for protein fold pattern recognition, Bioinformatics, 22 (2006) 1717-1722.
[58] L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review, 33 (2010) 1-39.
[59] D.J. Yu, J. Hu, Y. Huang, H.B. Shen, Y. Qi, Z.M. Tang, J.Y. Yang, TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble, Journal of Computational Chemistry, 34 (2013) 974-985.
[60] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica Et Biophysica Acta, 405 (1975) 442-451.

[61] B. Liu, iEnhancer-PsedeKNC: Identification of enhancers and their subgroups based on Pseudo degenerate kmer nucleotide composition, Neurocomputing, 217 (2016) 46-52.
[62] Z.Y. Liang, H.Y. Lai, H. Yang, C.J. Zhang, H.H. Wei, X.X. Chen, Y.W. Zhao, Z.D. Su, W.C. Li, Pro54DB: a database for experimentally verified sigma-54 promoters, Bioinformatics, 33 (2017) 467-469.
[63] H. Yang, H. Tang, X.X. Chen, C.J. Zhang, P.P. Zhu, H. Ding, W. Chen, H. Lin, Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition, BioMed Research International, 2016 (2016) 5413903.
[64] X.X. Chen, T. Hua, W.C. Li, W. Hao, C. Wei, D. Hui, L. Hao, Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition, BioMed Research International, 2016 (2016) 1654623.

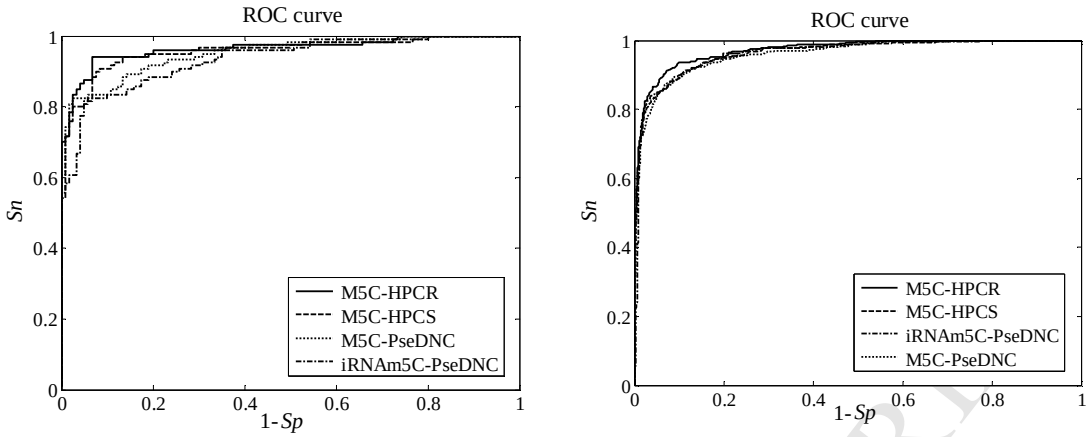


Fig. 4. The ROC curves of M5C-PseDNC, iRNAm5C-PseDNC, M5C-HPCS and the proposed M5C-HPCR on (A) Met1320 subset and (B) Met1900 dataset.