Contents lists available at ScienceDirect

Analytical Biochemistry

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

Improving N^6-methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties

Ming Zhang [a, b, **], Jia-Wei Sun [b], Zi Liu [a], Ming-Wu Ren [a], Hong-Bin Shen [c], Dong-Jun Yu [a, *]

[a] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[b] *School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China*
[c] *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China*

ARTICLE INFO

Article history:
Received 27 March 2016
Received in revised form
31 May 2016
Accepted 1 June 2016
Available online 11 June 2016

Keywords:
RNA sequence
N^6-methyladenosine
Physical–chemical property selection
Feature representation
Support vector machine

ABSTRACT

N^6-methyladenosine (m^6A) is one of the most common and abundant post-transcriptional RNA modifications found in viruses and most eukaryotes. m^6A plays an essential role in many vital biological processes to regulate gene expression. Because of its widespread distribution across the genomes, the identification of m^6A sites from RNA sequences is of significant importance for better understanding the regulatory mechanism of m^6A. Although progress has been achieved in m^6A site prediction, challenges remain. This article aims to further improve the performance of m^6A site prediction by introducing a new heuristic nucleotide physical–chemical property selection (HPCS) algorithm. The proposed HPCS algorithm can effectively extract an optimized subset of nucleotide physical–chemical properties under the prescribed feature representation for encoding an RNA sequence into a feature vector. We demonstrate the efficacy of the proposed HPCS algorithm under different feature representations, including pseudo dinucleotide composition (PseDNC), auto-covariance (AC), and cross-covariance (CC). Based on the proposed HPCS algorithm, we implemented an m^6A site predictor, called M6A-HPCS, which is freely available at http://csbio.njust.edu.cn/bioinf/M6A-HPCS. Experimental results over rigorous jackknife tests on benchmark datasets demonstrated that the proposed M6A-HPCS achieves higher success rates and outperforms existing state-of-the-art sequence-based m^6A site predictors.

N^6-methyladenosine (m^6A) is one of the most prevalent post-transcriptional RNA modifications [1–3] and plays a critical role in a number of biological processes, including mRNA splicing, export, stability, immune tolerance, and transcription [4–10]. Current research has revealed that the m^6A modification is a dynamic and reversible process that affects the gene regulation function in apoptosis, circadian rhythm, and meiosis [11–15]. Fig. S1 in online Supplementary Material S1 illustrates the reversible procedure of N^6-methylation and demethylation in mRNA. Besides that, m^6A modification is intimately correlated with human diseases, including obesity, cancer, infertility, and hepatitis [4,16–20]. Therefore, accurately obtaining knowledge of m^6A is vitally important for both basic research and drug development.

With the development of high-throughput m^6A profiling techniques such as MeRIP-Seq [2,21] and m6A-seq [22], the transcriptome-wide maps of m^6A distributions are now available for several species, including *Oryza sativa* [23], *Saccharomyces cerevisiae* [15], *Mus musculus* [1], and *Homo sapiens* [1]. These studies have provided deep insights into the distributions of m^6A modification sites and stimulated the development of this area. Nevertheless, the current pure biochemical experimental methods for targeting m^6A sites 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^6A sites from RNA sequences would be especially useful.

Abbreviations used: m^6A, N^6-methyladenosine; PseDNC, pseudo dinucleotide composition; AC, auto-covariance; CC, cross-covariance; SVM, support vector machine; HPCS, heuristic nucleotide physical–chemical property selection; PseAAC, pseudo amino acid composition; PCM, physical–chemical property matrix; RGoPCP, relative gain of physical–chemical property; Sp, specificity; Sn, sensitivity; Acc, accuracy; MCC, Matthews correlation coefficient; ROC, receiver operating characteristic; AUC, area under the ROC curve.

* Corresponding author.
** Corresponding author. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.
E-mail addresses: zhangming@just.edu.cn (M. Zhang), njyudj@njust.edu.cn (D.-J. Yu).

http://dx.doi.org/10.1016/j.ab.2016.06.001
0003-2697/© 2016 Elsevier Inc. All rights reserved.

Until now, several computational methods have been developed to predict post-translational modification sites, such as lysine ubiquitination sites [24], enzyme catalytic sites [25], lysine succinylation sites [26], and methylation sites [27–31], from primary sequences. Among these methods, Chen and coworkers performed pioneering work and developed a first m^{6}A site predictor called iRNA-Methyl [28]. Each RNA sequence sample in iRNA-Methyl was formulated with pseudo dinucleotide composition (PseDNC), into which three RNA physical–chemical properties were incorporated. However, the performance of iRNA-Methyl is not satisfactory, and the overall accuracy success rate is only 65.59%. Recently, Liu and coworkers [27] developed a machine-learning-based predictor called pRNAm-PC. pRNAm-PC encodes each RNA sequence sample into a feature vector by performing a series of auto-covariance (AC) and cross-covariance (CC) transformations on a physical–chemical matrix of nucleotides and uses a support vector machine (SVM) as a prediction engine. Experimental results demonstrated that pRNAm-PC achieved remarkably higher success rates with an overall accuracy success rate of 69.74%.

Although progress has been made in designing computational methods for targeting m^{6}A sites from RNA sequences, there are still several issues that deserve further investigation. First, knowing which physical–chemical properties of nucleotides are prominent for targeting m^{6}A sites is still a problem. In iRNA-Methyl [28], only three physical–chemical properties of nucleotides—enthalpy, entropy, and free energy—were used to encode each RNA segment into a PseDNC feature; while in pRNAm-PC [27], 10 physical–chemical properties of nucleotides were used to construct the combined feature based on AC and CC transformations, and the prediction performance has been remarkably improved. However, there are many other physical–chemical properties of nucleotides that can be used such as local structural properties [32], entropy and enthalpy [33], energy [34,35], nucleotide content [36], hydrophilicity [37], and elastic behavior [38]. Nevertheless, in both iRNA-Methyl [28] and pRNAm-PC [27], the authors only provided the prediction performances under several physical–chemical properties but did not explain why these properties were selected. Hence, how to measure the significance of these physical–chemical properties for the prediction of m^{6}A sites is still a critical problem worthy of further study.

Second, although pRNAm-PC performed much better than previous predictors, its prediction performances still need to be improved for potential practical application.

This article aims to address these two issues, and a new m^{6}A site predictor called M6A-HPCS (heuristic nucleotide physical–chemical property selection) is proposed. First, the concepts of the *relative gain* and *direct gain* of physical–chemical property are introduced to measure the significance of a physical–chemical property for targeting the m^{6}A sites. Then, a heuristic algorithm based on the *relative gains* and *direct gains* of physical–chemical properties is put forward to optimize a subset of physical–chemical properties under the prescribed feature representation method. After that, RNA samples can be encoded into feature vectors according to the optimized subset of physical–chemical properties. Finally, an SVM is used as a prediction engine to construct M6A-HPCS on the training dataset with the developed feature set. Rigorous jackknife tests on the benchmark dataset demonstrate the superiority of the proposed M6A-HPCS over the existing m^{6}A site predictors. Below, we describe the aforementioned steps one by one.

1. Materials and methods

1.1. Benchmark dataset

In this study, the datasets were constructed from 1183 genes in *S. cerevisiae* genome [15]. Wet lab experiments identified that the genome contained 1307 methylated adenine sites and 33,280 non-methylated adenine sites with a consensus motif “GAC” [15]. To obtain a high-quality training dataset with balanced class samples, 1307 negative samples from the 33,280 non-methylated adenine sites were selected as negative samples [28]. The 1307 positive and 1307 negative samples constitute the final dataset. Note that the maximal sequence identity among the dataset was reduced to 85% by using CD-HIT [39], and the self-conflicted samples in the dataset were also excluded [28].

The RNA samples in the dataset can be uniformly formulated as

$$R_\xi(\mathrm{GAC}) = N_{-\xi}N_{-(\xi-1)}\cdots N_{-2}N_{-1}\mathrm{GAC}N_{+1}N_{+2}\cdots N_{+(\xi-1)}N_{+\xi}, \tag{1}$$

where the GAC represents the consensus motif, the center A represents adenine, the subscript ξ is an integer, $N_{-\xi}$ represents the ξ-th upstream nucleotide from the center GAC, and $N_{+\xi}$ represents the ξ-th downstream nucleotide from the center GAC. Hence, the length of RNA sequence segment R_ξ(GAC) is $2\xi+3$. Previous studies [28,40] have proven that ξ=24 is a better choice for designing a machine-learning-based m^{6}A site predictor. In this study, we also took this configuration. Accordingly, each RNA sequence sample in the benchmark dataset consists of $(2\xi+3)=51$ nucleotides. Each R_ξ(GAC) can be further classified into the following two categories:

$$R_\xi(\mathrm{GAC}) \in \begin{cases} R_\xi^+(\mathrm{GAC}), & \text{if its centre is a methylation site} \\ R_\xi^-(\mathrm{GAC}), & \text{otherwise} \end{cases}, \tag{2}$$

where R_ξ^+(GAC) denotes a positive sequence sample if its center adenine can be N^6-methylated, whereas R_ξ^-(GAC) denotes a negative sequence sample if its center adenine cannot be N^6-methylated, the symbol $\in$ means “a member of” in set theory. Hence, the benchmark dataset S_ξ can be formulated as

$$S_\xi = S_\xi^+ \cup S_\xi^-, \tag{3}$$

where the positive subset S_ξ^+ consists only of the methylation RNA segments, S_ξ^- contains only the samples of the non-methylation RNA segments, and $\cup$ represents the symbol for “union” in set theory.

1.2. Feature representation of RNA sequence

Each RNA sequence sample, denoted as $\boldsymbol{R}$, in the dataset as defined in Eq. (3) can be reformulated as follows:

$$\boldsymbol{R} = N_1N_2\cdots N_i\cdots N_L, \tag{4}$$

where N_i $(1 \le i \le L)$ represents the i-th nucleotide in RNA sequence sample $\boldsymbol{R}$ and L is the length of $\boldsymbol{R}$. Each N_i belongs to one of the four native nucleotides, that is, $N_i \in$ {A (adenine), C (cytosine), G (guanine), U (uracil)}. For designing a machine-learning-based m^{6}A site predictor, a critical step is how to transform each RNA sequence sample as formulated in Eq. (4) to a feature vector with fixed length. The underlying reason is that most of the existing machine-learning algorithms can handle only vectors but not sequence samples, as elaborated in Ref. [41]. Motivated by the wide and successful use of pseudo amino acid composition or Chou's pseudo amino acid composition (PseAAC) in the areas of computational proteomics with protein/peptide sequences [42,43], recently the concept of pseudo *k*-tuber nucleotide composition has been developed to deal with DNA/RNA sequences in computational genetics and genomics [44–48]. According to Ref. [49], the general form of pseudo composition for RNA sequence samples can be formulated by

$$[\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_i \quad \cdots \quad \Psi_\Omega]^T, \tag{5}$$

where the symbol T is the transpose operator and the subscript Ω is an integer to reflect the vector's dimension. The value of Ω, as well as the components Ψ_i $(i = 1,2, \cdots, \Omega)$, will depend on how to extract the desired information from the RNA sequence sample of Eq. (4). Next, we describe how to encode each RNA sequence sample into a fixed-length feature vector as formulated in Eq. (5) based on the so-called physical–chemical property matrix.

1.2.1. Physical–chemical property matrix

Because two individual nucleotides can be polymerized into a dimer or dinucleotide, there are in total $4 \times 4 = 16$ types of native dinucleotides in RNA sequences: AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU. Clearly, different dinucleotides possess different physical–chemical properties. Hence, we can encode an RNA sequence into a feature vector by using multiple physical–chemical properties. In the current study, the following 23 types of physical–chemical (PC) properties were considered: (1) PC^1: rise [32]; (2) PC^2: roll [32]; (3) PC^3: shift [32]; (4) PC^4: slide [32]; (5) PC^5: tilt [32]; (6) PC^6: twist [32]; (7) PC^7: stacking energy [32]; (8) PC^8: enthalpy [33]; (9) PC^9: enthalpy2 [33]; (10) PC^{10}: entropy [33]; (11) PC^{11}: entropy2 [34]; (12) PC^{12}: free energy [34]; (13) PC^{13}: free energy2 [34]; (14) PC^{14}: adenine content [36]; (15) PC^{15}: cytosine content [36]; (16) PC^{16}: GC content [36]; (17) PC^{17}: guanine content [36]; (18) PC^{18}: keto (GT) content [36]; (19) PC^{19}: purine (AG) content [36]; (20) PC^{20}: thymine content [36]; (21) PC^{21}: hydrophilicity [37]; (22) PC^{22}: hydrophilicity2 [37]; (23) PC^{23}: base stacking energy [35]. Table S1 in online Supplementary Material S2 summarizes the original values of the 23 physical–chemical properties.

To facilitate the subsequent computation, we first normalize the values of the 23 physical–chemical properties in Table S1 using the following equation:

$$\mathrm{PC}^i_{normalized}(j) = \frac{\mathrm{PC}^i(j) - Mean(i)}{Std(i)}, \tag{6}$$

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

Let u be the number of physical–chemical properties considered ($u = 23$ in this study). Then, based on the normalized values of physical–chemical properties, an RNA sample $\boldsymbol{R}$ with L nucleotides can be formulated with a $u \times (L{-}1)$ physical–chemical property matrix (PCM), denoted as $\mathrm{PCM}(\boldsymbol{R})$, as follows:

$$\mathrm{PCM}(\boldsymbol{R}) = \begin{bmatrix} \mathrm{PC}^1(N_1N_2) & \mathrm{PC}^1(N_2N_3) & \cdots & \mathrm{PC}^1(N_{L-1}N_L) \\ \mathrm{PC}^2(N_1N_2) & \mathrm{PC}^2(N_2N_3) & \cdots & \mathrm{PC}^2(N_{L-1}N_L) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{PC}^u(N_1N_2) & \mathrm{PC}^u(N_2N_3) & \cdots & \mathrm{PC}^u(N_{L-1}N_L) \end{bmatrix}, \tag{7}$$

where $\mathrm{PC}^i(N_jN_{j+1})$ is the i-th $(1 \le i \le u)$ physical–chemical property value for the N_jN_{j+1} $(1 \le j \le L{-}1)$ dinucleotide in R.

The next key problem is how to extract effective feature for m^6A site prediction from the PCM of a given RNA sequence sample. Currently, to the best of our knowledge, only two types of features—PseDNC [28] and the combination of AC and CC [27]—have been investigated for m^6A site prediction, both based on PCM. In view of this, we consider these two types of features to demonstrate the effectiveness of the proposed physical–chemical property selection algorithm (refer to "Optimized subsets of physical–chemical properties" section in Results and Discussion below).

1.2.2. PseDNC feature

Encouraged and stimulated by the successes of PseAAC [42,43] in dealing with protein/peptide sequence, the concept of PseDNC has been proposed to represent DNA/RNA sequence for identifying m^6A sites [28]. Given an RNA sequence sample $\boldsymbol{R}$ with L nucleotides as defined in Eq. (4), the PseDNC feature vector of $\boldsymbol{R}$, denoted as $\mathbf{f}_{\mathrm{PseDNC}}$, can be formulated as follows [28]:

$$\mathbf{f}_{\mathrm{PseDNC}} = [f_1 \quad f_2 \quad \cdots \quad f_{16} \quad f_{16+1} \quad f_{16+2} \quad \cdots \quad f_{16+\lambda}]^T, \tag{8}$$

where

$$f_k = \begin{cases} \dfrac{d_k}{\sum_{i=1}^{16} d_i + w\sum\limits_{j=1}^{\lambda}\theta_i} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} d_i + w\sum\limits_{j=1}^{\lambda}\theta_i} & (16 < k \le 16+\lambda) \end{cases} \tag{9}$$

where d_k $(1 \le k \le 16)$ is the normalized occurrence frequency of the k-th non-overlapping dinucleotides in the RNA sequence and θ_j $(1 \le j \le \lambda)$ is the j-tier correlation factor, which reflects the sequence order correlation between all of the most contiguous dinucleotides along the RNA sequence sample, defined as follows:

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Psi(N_iN_{i+1}, N_{i+j}N_{i+j+1}) \quad (j = 1,\ 2,\ \cdots,\ \lambda;\ \lambda < L-1), \tag{10}$$

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

$$\Psi(N_iN_{i+1}, N_{i+j}N_{i+j+1}) = \frac{1}{\mu}\sum_{t=1}^{\mu}\left[\mathrm{PC}^t(N_iN_{i+1}) - \mathrm{PC}^t(N_{i+j}N_{i+j+1})\right]^2, \tag{11}$$

where μ is the number of the physical–chemical properties considered and $\mathrm{PC}^t(N_iN_{i+1})$ represents the value of the t-th physical–chemical property for the dinucleotides N_iN_{i+1} in the RNA sequence.

In Eq. (9), the parameter λ is an integer representing the highest counted tier of the correlation along the RNA sequence, whereas w is the weight factor ranging from 0 to 1 for balancing the significance of 2-tuple nucleotide compositions and correlation factors. Accordingly, the dimensionality of the PseDNC feature is $16 + \lambda$.

1.2.3. Features derived from AC and CC transformations

Recently, Liu and coworkers [27] proposed a new and more discriminative feature representation method for m^6A site prediction. This method extracts the feature of an RNA sequence by performing a series of AC and CC transformations with multiple physical–chemical properties on the PCM of the RNA sequence.

For a given RNA sequence sample $\boldsymbol{R}$ with L nucleotides, a scalar quantity AC (i,g), which reflects the correlation of the i-th physical–chemical property between two subsequences separated by g dinucleotides, can be extracted by performing AC transformation on the i-th row of PCM($\boldsymbol{R}$) (refer to Eq. (7)) as follows:

$$\mathrm{AC}(i,g)=\frac{\sum_{j=1}^{L-1-g}\left(\mathrm{PC}^i(N_jN_{j+1})-\overline{\mathrm{PC}^i}\right)\left(\mathrm{PC}^i(N_{j+g}N_{j+1+g})-\overline{\mathrm{PC}^i}\right)}{L-1-g}, \tag{12}$$

where $1\le g\le G$, $1\le G\le L{-}2$, and $1\le i\le u$. Note that G is a predefined integer and is the maximum value of g, $\underline{u \text{ is}}$ the number of physical–chemical properties considered, and $\overline{\mathrm{PC}^i}$ represents the mean of the values in the i-th row of PCM($\boldsymbol{R}$) as defined in Eq. (7), as given by

$$\overline{\mathrm{PC}^i}=\frac{\sum_{j=1}^{L-1}\mathrm{PC}^i(N_jN_{j+1})}{L-1}. \tag{13}$$

Then, the set of AC feature components extracted by using Eq. (12) can be formulated as follows:

$$\{\mathrm{AC}(i,g)|1\le i\le u,\ \ 1\le g\le G\}. \tag{14}$$

According to Eq. (14), we can clearly find that the dimensionality of the AC feature is $u\times G$ if u types of physical–chemical properties are considered.

Similarly, a scalar quantity CC (i_1,i_2,g) reflects that the correlation between two subsequences, each belonging to different physical–chemical properties, can be extracted by performing CC transformation on PCM($\boldsymbol{R}$) (refer to Eq. (7)) as follows:

$$\mathrm{CC}(i_1,i_2,g)=\frac{\sum_{j=1}^{L-1-g}\left(\mathrm{PC}^{i_1}(N_jN_{j+1})-\overline{\mathrm{PC}^{i_1}}\right)\left(\mathrm{PC}^{i_2}(N_{j+g}N_{j+1+g})-\overline{\mathrm{PC}^{i_2}}\right)}{L-1-g}, \tag{15}$$

where $1\le i_1\le u, 1\le i_2\le u, i_1\ne i_2, 1\le g\le G$, and $1\le G\le L{-}2$, and G is a predefined integer. Then, the set of CC feature components extracted by using Eq. (15) can be formulated as follows:

$$\{\mathrm{CC}(i_1,i_2,g)|1\le i_1\le u,\quad 1\le i_2\le u,\quad i_1\ne i_2,\quad 1\le g\le G\}. \tag{16}$$

Accordingly, the dimensionality of the CC feature extracted by using Eq. (16) is $u\times(u{-}1)\times G$ if u types of physical–chemical properties are considered. Therefore, the dimensionality of the combination of AC and CC feature vector, denoted as AC + CC, is $u\times G+u\times(u{-}1)\times G=u\times u\times G$.

1.3. Optimize the subset of physical–chemical properties using a heuristic algorithm

As described in the "Feature representation of RNA sequence" section above, currently only two types of features—PseDNC [28] and the combination of AC and CC [27]—have been investigated for m^{6}A site prediction. On the other hand, the discriminative capability of both features heavily depends on the physical–chemical properties used. In Ref. [28] three physical–chemical properties of nucleotides—enthalpy, entropy, and free energy—were used to extract the PseDNC feature, whereas in Ref. [27] ten physical–chemical properties of nucleotides were taken to construct the combination of AC and CC features. However, none of them explains why those properties were selected from the physical–chemical properties set. It would be especially useful if we could distinguish which physical–chemical properties will have much more positive impacts toward the prediction of m^{6}A sites. In the subsequent sections, we aimed to solve this problem by using a heuristic algorithm that can select multiple optimized subsets of physical–chemical properties based on their significances.

1.3.1. Measure the relative gain of physical–chemical property

Let $S_{all}=\{\mathrm{PC}^i\}_{i=1}^{u}$ be the set of u known physical–chemical properties and S be a subset of S_{all}. Clearly, we can use physical–chemical property-related feature representation methods (e.g., PseDNC or AC + CC described above) based on the elements in S to encode RNA sequence samples. For the convenience of subsequent description, we uniformly term these feature representation methods, which extract features based on S, as $\mathbf{f}(S)$.

We define the relative gain of physical–chemical property (RGoPCP) as follows:

$$rg\left(\mathrm{PC}^i,\mathbf{f},S\right)=\mathrm{Acc}\left(\mathbf{f}\left(S\cup\left\{\mathrm{PC}^i\right\}\right)\right)-\mathrm{Acc}(\mathbf{f}(S)), \tag{17}$$

where PC^i ($\mathrm{PC}^i\in S_{all}$ and $\mathrm{PC}^i\notin S$) represents the i-th physical–chemical property and Acc(S) and Acc($\mathbf{f}(S\cup\{\mathrm{PC}^i\})$) represent the overall prediction accuracies under the feature representations $\mathbf{f}(S)$ and $\mathbf{f}(S\cup\{\mathrm{PC}^i\})$, respectively, with a prescribed prediction engine over cross-validation on a given dataset.

Note that the overall prediction accuracy (i.e., Acc in Eq. (17)) can be evaluated by any feasible types of cross-validation. In this study, 10-fold cross-validation was taken. As to prediction engine, any machine-learning algorithm can be used. In this study, we took SVM as the prediction engine.

According to Eq. (17), the RGoPCP measures the *relative gain* of accuracy for a physical–chemical property over a given subset of physical–chemical properties. $rg\,(\mathrm{PC}^i,\mathbf{f},S)>0$ represents a positive gain; that is, the prediction accuracy generated by the physical–chemical property subset $S\cup\{\mathrm{PC}^i\}$ is higher than that generated by the subset S. Accordingly, $rg(\mathrm{PC}^i,\mathbf{f},S)<0$ denotes a negative gain, where adding PC^i into S will deteriorate the prediction accuracy.

We define $\mathrm{Acc}(S)=0$ when $S=\Phi$ (i.e., no physical–chemical property exists in S). Then, Eq. (17) can be rewritten as

$$rg\left(\mathrm{PC}^i,\mathbf{f},\Phi\right)=\mathrm{Acc}\left(\mathbf{f}\left(\left\{\mathrm{PC}^i\right\}\right)\right). \tag{18}$$

In other words, $rg(\mathrm{PC}^i,\mathbf{f},\Phi)$ defined in Eq. (18) measures the *direct gain* of the individual physical–chemical property PC^i. In the subsequent section, we demonstrate how to optimize the selection of physical–chemical properties for m^{6}A site prediction based on the *relative gain* and *direct gain* defined in Eqs. (17) and (18), respectively.

1.3.2. Optimize the subset of physical–chemical properties

Finding the optimal subset from a given physical–chemical property set for a specific prediction task (e.g., m^{6}A site prediction) is in fact a combinatorial explosion problem. The time complexity of a brute-force algorithm for selecting the optimal subset is $O(2^u)$, where u is the number of physical–chemical properties considered. In view of this, here we present a heuristic algorithm, denoted as HPCS, which can select a suboptimal subset from the given physical–chemical property set with lower time complexity.

Let $S_{all} = \{\mathrm{PC}^i\}_{i=1}^{u}$ be the set of u physical–chemical properties, S be a subset of S_{all}, $\mathbf{f}(S)$ be the feature representation method based on S, and K be the predefined positive integer denoting how many candidate subsets are generated from S_{all}.

The proposed heuristic algorithm first ranks all of the physical–chemical properties in S_{all} according to their *direct gains* in descending order. Then, the top K physical–chemical properties are used to construct K initial candidate subsets. After that, the K initial candidate subsets are gradually expanded by considering the *relative gains* of physical–chemical properties. Finally, among the K candidate subsets, the one that can achieve the best Acc is chosen as the optimized subset. We describe the details of the proposed heuristic 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 *direct gain* using Eq. (18). Then, the u physical–chemical properties are ranked according to their *direct gains* in descending order, denoted as

$$S_{Rank} = \left\{\mathrm{PC}^k_{Rank}\right\}_{k=1}^{u}.$$

Step 2: Initialize the K candidate physical–chemical subsets.

The top K physical–chemical properties are selected to initialize the K candidate subsets as follows:

$$S_k \leftarrow \left\{\mathrm{PC}^k_{Rank}\right\};\ 1 \le k \le K. \tag{19}$$

Step 3: Expand the K candidate physical–chemical subsets.

Each of the K initial subsets will be gradually expanded by considering the *relative gains* of the remaining elements in S_{Rank}.

Taking the k-th initial subset as an example, we expand it with an iterative procedure as follows.

For each of the remaining elements in S_{Rank}–S_k, denoted as PC^j_{Rank}, we first compute its *relative gain*—that is, $rg(\mathrm{PC}^j,\mathbf{f},S_k)$—according to Eq. (17).

Then, we can locate the j^*-th element, which has the maximal value of rg, from S_{Rank}–S_k as follows:

$$j^* = \max_j rg\left(\mathrm{PC}^j, \mathbf{f}, S_k\right),\ \text{where } \mathrm{PC}^j \in S_{Rank} - S_k. \tag{20}$$

If $rg(\mathrm{PC}^{j^*}, \mathbf{f}, S_k) > 0$, the j^*-th element will be added into S_k (Eq. (21)) because it still has positive *relative gain* over S_k:

$$S_k \leftarrow S_k \cup \left\{\mathrm{PC}^{j^*}\right\}. \tag{21}$$

This expansion process for the k-th subset continues until $rg(\mathrm{PC}^{j^*}, \mathbf{f}, S_k) \le 0$ (i.e., no positive *relative gain* can be made).

Step 4: Choose the best one from K candidate subsets as the optimized subset.

After the K candidate subsets have been identified, the one that can achieve the highest value of Acc is chosen as the optimized subset, denoted as $S_{optimized\_subset}$:

$$k^* = \arg\max_{1 \le k \le K} \mathrm{Acc}(\mathbf{f}(S_k)). \tag{22}$$

$$S_{optimized\_subset} \leftarrow S_{k^*}. \tag{23}$$

Algorithm 1 summarizes the procedure of HPCS for optimizing the subset of physical–chemical properties based on RGoPCP. Note that the parameter K, which is a positive integer ($1 < K \le u$, where u is the number of physical–chemical properties considered) denoting how many candidate subsets are generated from S_{all}, needs to be prescribed before executing the algorithm. Clearly, more candidate subsets could be generated with a larger value of K and, thus, the probability of obtaining the optimal subset will be higher. However, the computational complexity will also be high with a large value of K. We should make a trade-off between the performance and the computational efficiency. We have tested different values of K ($K = 3$, $K = 5$, $K = 7$, $K = 9$, and $K = 11$) and found that the optimal subset obtained by the algorithm is the same when $K \ge 5$. In view of this, we set $K = 5$ in this study.

As to computational efficiency, it is easy to calculate that the

Algorithm 1
HPCS: A heuristic algorithm for optimizing subset of physical–chemical properties.

```
Input:  $S_{all}\{\mathrm{PC}^i\}_{i=1}^{u}$: set of u physical–chemical properties; $\mathbf{f}(\cdot)$:
        predefined feature representation method; K: predefined
        positive integer denoting how many subsets are generated
Output: $S_{optimized\_subset}$: Optimized physical–chemical subset of $S_{all}$
Step 1       Rank the u physical–chemical properties
        1.1  For each of the physical–chemical properties in $S_{all}$,
             calculate its direct gain using Eq. (18) as follows:
             $rg(\mathrm{PC}^i, \mathbf{f}, \Phi) = \mathrm{Acc}(\mathbf{f}(\{\mathrm{PC}^i\}))$
        1.2  Rank the u physical–chemical properties according to
             their direct gains in descending order:
             $S_{Rank} = \{\mathrm{PC}^k_{Rank}\}_{k=1}^{u}$
Step 2       Initialize the K candidate physical–chemical subsets
        2.1  FOR $k = 1$, L, K
        2.2  $S_k \leftarrow \{\mathrm{PC}^k_{Rank}\}$
        2.3  END FOR
Step 3       Expand the K candidate physical–chemical subsets
        3.1  FOR $k = 1$, L, K
        3.2  WHILE (TRUE)
        3.3  For each of the remaining elements in $S_{Rank}$–$S_k$, denoted
             as $\mathrm{PC}^j_{Rank}$, compute its relative gain [i.e., $rg(\mathrm{PC}^j, \mathbf{f}, S_k)$]
             according to Eq. (17)
        3.4  Locate the $j^*$-th element, which has the maximal value of
             rg, from $S_{Rank}$–$S_k$ as follows:
             $j^* = \max_j rg(\mathrm{PC}^j, \mathbf{f}, S_k)$, where $\mathrm{PC}^j \in S_{Rank} - S_k$
        3.5  IF $rg(\mathrm{PC}^{j^*}, \mathbf{f}, S_k) > 0$
        3.6  $S_k \leftarrow S_k \cup \{\mathrm{PC}^{j^*}\}$
        3.7  ELSE
        3.8  BREAK WHILE
        3.9  END IF
        3.10 END WHILE
        3.11 END FOR
Step 4       Choose the best one from K candidate subsets as the
             optimized subset
        4.1  Among the K candidate subsets, the one that achieves
             the highest value of Acc is chosen as the optimized one:
             $k^* = \arg\max_{1 \le k \le K} \mathrm{Acc}(\mathbf{f}(S_k)) S_{optimized\_subset} \leftarrow S_{k^*}$
        4.2  RETURN $S_{optimized\_subset}$
```

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.

1.4. SVM classifier

Support vector machine, which was proposed by Cortes and Vapnik [50], has been widely used in the realm of bioinformatics [27–29,46,48,51–53]. The basic idea of SVM is to transform the input vector into a high-dimension Hilbert space by kernel functions and then seek a separating hyper plane between classes with the maximal margin in this space. For more information about SVM, refer to Refs. [54–56].

In this study, the LIBSVM package [57,58], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm, was taken to implement an SVM classifier. The popular radial basis function (RBF) was chosen as the kernel function, where the regularization parameter C and the kernel width parameter γ were optimized based on 10-fold cross-validation using a grid search strategy in the LIBSVM package.

1.5. Performance metrics

Four routinely used indexes in this field—specificity (Sp), sensitivity (Sn), accuracy (Acc), and the Matthews correlation coefficient (MCC) [59]—were taken to evaluate the prediction performances as follows:

$$\begin{cases} \mathrm{Sp} = 1\text{-}\dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Acc} = 1\text{-}\dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \tag{24}$$

where N^+ is the total number of positive samples or true methylation RNA sequence investigated, N^- is the total number of negative samples or non-methylation RNA sequence investigated, N_-^+ is the total number of true methylation RNA samples incorrectly predicted to be non-methylation RNA samples, and N_+^- is the total number of non-methylation RNA samples incorrectly predicted to be true.

In addition, the graph of receiver operating characteristic (ROC) and the area under the ROC curve (AUC) were used to evaluate the overall prediction qualities of the considered prediction models. The AUC is threshold independent and increases in direct proportion to prediction performance.

2. Results and discussion

Independent dataset test, sub-sampling (or K-fold cross-validation) test, and jackknife test (or leave-one-out cross-validation) have been the routinely used methods for evaluating the performances of statistical prediction models [60]. Because the jackknife test can always yield a unique outcome for a given benchmark dataset, it is considered to be the most objective and least arbitrary method. Therefore, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors [24,26–29,44,48,49,61–64].

In this study, the jackknife test was used to evaluate the performance of the proposed predictor. During the jackknife test, as elucidated in Ref. [49], each of the samples in the benchmark dataset is in turn singled out as an independent test sample and all of the remaining samples are used as training samples. However, the complexity of the jackknife test is equal to the data volume in the dataset, which makes it time-consuming to implement. In view of this, in this study 10-fold cross-validation was used to accelerate those applications that comprise multiple iterative procedures (e.g., SVM parameter optimization and physical–chemical property subset optimization in HPCS algorithm), whereas the rigorous jackknife test was adopted to comprehensively evaluate performances of different m^{6}A site predictors.

2.1. Results of direct gain of physical–chemical property

We calculated the *direct gains* of the 23 physical–chemical properties with Eq. (18) under both PseDNC and AC + CC feature representations. Tables 1 and 2 list the ranked physical–chemical properties according to their *direct gains* in descending order under PseDNC and AC + CC as feature representations, respectively.

From Table 1, the partial order relationship among the 23 physical–chemical properties under PseDNC feature representation can be formulated according to their *direct gains* as follows:

$$\begin{gathered} PC^2 \succeq PC^3 \succeq PC^{11} \succeq PC^9 \succeq PC^{22} \succeq PC^{12} \succeq PC^{19} \succeq PC^{10} \\ \succeq PC^{16} \succeq PC^{13} \succeq PC^{21} \succeq \\ PC^8 \succeq PC^{14} \succeq PC^4 \succeq PC^{18} \succeq PC^{23} \succeq PC^5 \succeq PC^{20} \succeq PC^1 \\ \succeq PC^{15} \succeq PC^6 \succeq PC^{17} \succeq PC^7, \end{gathered} \tag{25}$$

where the symbol "≥" is the partial order operator, which means "greater than or equal to."

Similarly, the partial order relationship among the 23 physical–chemical properties under AC + CC feature representation can be formulated according to their *direct gains* listed in Table 2 as follows:

$$\begin{gathered} PC^{11} \succeq PC^9 \succeq PC^3 \succeq PC^8 \succeq PC^{10} \succeq PC^{21} \succeq PC^{12} \\ \succeq PC^{18} \succeq PC^{15} \succeq PC^{23} \succeq PC^{13} \succeq \\ PC^{14} \succeq PC^{16} \succeq PC^{22} \succeq PC^2 \succeq PC^{19} \succeq PC^{17} \succeq PC^4 \\ \succeq PC^{20} \succeq PC^5 \succeq PC^7 \succeq PC^1 \succeq PC^6. \end{gathered} \tag{26}$$

By analyzing Tables 1 and 2, together with Eqs. (25) and (26), several observations can be made. First, the *direct gain* of a given physical–chemical property is closely related to the feature representation. As shown in Table 1, the *direct gains* of the 23 physical–chemical properties under PseDNC feature representation are all larger than 63%, whereas those under AC + CC feature

Table 1
Ranked direct gains of the 23 physical–chemical properties in descending order under PseDNC feature representation.

Rank	Properties	Direct gain (%)	Rank	Properties	Direct gain (%)
1	PC^2	66.72	13	PC^{14}	65.30
2	PC^3	66.49	14	PC^4	65.26
3	PC^{11}	66.03	15	PC^{18}	65.07
4	PC^9	65.99	16	PC^{23}	64.88
5	PC^{22}	65.80	17	PC^5	64.80
6	PC^{12}	65.76	18	PC^{20}	64.80
7	PC^{19}	65.68	19	PC^1	64.65
8	PC^{10}	65.53	20	PC^{15}	64.42
9	PC^{16}	65.53	21	PC^6	64.31
10	PC^{13}	65.49	22	PC^{17}	64.00
11	PC^{21}	65.49	23	PC^7	63.96
12	PC^8	65.46			

Table 2
Ranked direct gains of the 23 physical–chemical properties in descending order under AC + CC feature representation.

Rank	Properties	Direct gain (%)	Rank	Properties	Direct gain (%)
1	PC^{11}	57.84	13	PC^{16}	55.01
2	PC^{9}	57.50	14	PC^{22}	55.01
3	PC^{3}	57.27	15	PC^{2}	54.55
4	PC^{8}	57.04	16	PC^{19}	54.55
5	PC^{10}	56.85	17	PC^{17}	54.51
6	PC^{21}	56.73	18	PC^{4}	54.44
7	PC^{12}	56.31	19	PC^{20}	54.44
8	PC^{18}	56.31	20	PC^{5}	54.17
9	PC^{15}	55.93	21	PC^{7}	52.83
10	PC^{23}	55.89	22	PC^{1}	52.64
11	PC^{13}	55.85	23	PC^{6}	52.30
12	PC^{14}	55.62			

representation are all smaller than 58%, as illustrated in Table 2.

Second, the partial order relationship of physical–chemical properties also heavily depends on the feature representation. As shown in Eqs. (25) and (26), different partial order relationships were extracted under PseDNC and AC + CC feature representations.

Third, we find that there does exist certain common-ness between the two different partial order relationships (refer to Eqs. (25) and (26)), although they are heavily affected by feature representations. For example, 7 physical–chemical properties—PC^3, PC^{11}, PC^9, PC^{12}, PC^{10}, PC^{13}, and PC^{21}—are ranked among the top 11 properties in both PseDNC- and AC + CC-based partial order relationships. Similarly, 7 common properties—PC^4, PC^5, PC^{20}, PC^1, PC^6, PC^{17}, and PC^7—appear in the last 11 properties in both partial order relationships.

2.2. Optimized subsets of physical–chemical properties

In this section, we apply the proposed heuristic Algorithm 1 to extract optimized subset of physical–chemical properties for m^{6}A site prediction. We executed Algorithm 1 under PseDNC and AC + CC feature representations, respectively. Note that we set the value of parameter K to be 5, and the 10-fold cross-validation was used to evaluate the Acc in applying Algorithm 1. Tables 3 and 4 illustrate the 5 generated candidate subsets of physical–chemical properties under PseDNC and AC + CC feature representations, respectively, after applying Algorithm 1, whereas Fig. 1 plots the comparisons of prediction accuracies (Acc) among the 5 candidate subsets under each feature representation. Among the 5 generated candidate subsets under each feature representation, the one that achieves the highest value of Acc is selected as the final optimized subset, as highlighted in bold in Tables 3 and 4. Note that S_{all} in Tables 3 and 4 denotes the full set of 23 physical–chemical properties.

By carefully analyzing the results illustrated in Tables 3 and 4, and Fig. 1, several observations can be made. First, there are two optimized subsets—S_2 and S_3—under PseDNC feature representation because they achieve the same highest value of Acc (67.48%). We found that the proposed algorithm selected the same 5 physical–chemical properties: {PC^3, PC^{11}, PC^{19}, PC^9, PC^6} = {shift, entropy2, purine (AG) content, enthalpy2, twist}, for S_2 and S_3. The only difference between S_2 and S_3 is the order in which the 5 physical–chemical properties are selected.

Table 3
The 5 generated candidate subsets and the full set of 23 physical–chemical properties under PseDNC feature representation.

Candidate subset	Selected physical–chemical properties	Acc (%)
S_1	PC^2	66.72
S_2	PC^3,PC^{11},PC^{19},PC^9,PC^6	67.48
S_3	PC^{11},PC^3, PC^{19},PC^9,PC^6	67.48
S_4	PC^9, PC^3, PC^{15}	66.79
S_5	PC^{22}, PC^{23}	66.37
S_{all}	All 23 physical–chemical properties	66.17

Second, we again found that the optimized subset is the second candidate subset (i.e., S_2) under AC + CC feature representation with the highest value of Acc (72.23%). Quite different from the PseDNC feature, under which only 5 of 23 physical–chemical properties were selected to construct the optimized subset, AC + CC feature representation produced 13 properties for constructing the optimized subset S_2 = {PC^9, PC^3, PC^{12}, PC^{14}, PC^6, PC^{18}, PC^4, PC^2, PC^7, PC^{23}, PC^{11}, PC^{10}, PC^8} = {enthalpy, shift, free energy, adenine content, twist, keto (GT) content, slide, roll, stacking energy, base stacking energy, entropy2, entropy, enthalpy}. Nevertheless, 4 of the 5 properties—PC^3, PC^{11}, PC^9, and PC^6—in the optimized subset under PseDNC feature representation also appear in the optimized subset under AC + CC feature representation, denoting that these 4 physical–chemical properties are especially useful for encoding RNA sequence for m^{6}A site prediction even under different feature representations.

Third, it is found that the performance of the optimized subset is consistently better than that of the full set of 23 physical–chemical properties under both PseDNC and AC + CC feature representation. In addition, the other 4 candidate subsets also outperform the full set under each of the two feature representations. These observations demonstrate that the proposed algorithm can really uncover those most important physical–chemical properties for m^{6}A site prediction.

2.3. Comparisons with existing m^{6}A site predictors

In this section, we compare the proposed method with several popular m^{6}A site predictors. For the purpose of fair comparison, we should compare the proposed method with other m^{6}A site predictors under the same feature representation. Because currently PseDNC and AC + CC are two major feature representations, both of which are physical–chemical property dependent, in existing m^{6}A site predictors we perform comparisons under these two feature representations over rigorous jackknife tests.

For the convenience of subsequent description, we term the proposed method as M6A-HPCS, which encodes RNA samples under either PseDNC or AC + CC feature representation with the proposed physical–chemical property selection procedure and takes SVM as the prediction engine. We also implemented SVM-based predictor, termed as M6A-SVM, under either PseDNC or AC + CC feature representation but without a physical–chemical property selection procedure.

2.3.1. Comparisons with predictors under PseDNC feature representation

Table 5 summarizes the comparison results among iRNA-Methyl [28], M6A-SVM, and M6A-HPCS, all of which are under PseDNC feature representation and use SVM as the prediction engine. Fig. 2A plots the ROC curves of the three predictors under PseDNC feature representation. Note that the differences among the three predictors are the numbers of nucleotide physical–chemical properties used to encode RNA samples to PseDNC features; iRNA-Methyl [28] used 3 physical–chemical properties of nucleotides (PC^8 [enthalpy], PC^{10} [entropy], and PC^{12} [free energy]), M6A-SVM took the entire set of 23 physical–chemical properties, and M6A-HPCS used an optimized physical–chemical subset ({PC^3, PC^{11}, PC^{19}, PC^9, PC^6} = {shift, entropy2, purine (AG) content, enthalpy2, twist}) of the 23 properties. Another issue is how to set the values of the two parameters—λ and w—in the PseDNC encoding scheme.

Table 4
The 5 generated candidate subsets and the full set of 23 physical–chemical properties under AC + CC feature representation.

Candidate subset	Selected physical–chemical properties	Acc (%)
S_1	PC^{11}, PC^{3}, PC^{13},PC^{14}, PC^{4}, PC^{21}, PC^{18}, PC^{19}, PC^{9}, PC^{22}, PC^{6}, PC^{15}	71.46
S_2	PC^{9}, PC^{3}, PC^{12}, PC^{14}, PC^{6}, PC^{18}, PC^{4}, PC^{2}, PC^{7}, PC^{23}, PC^{11}, PC^{10}, PC^{8}	72.23
S_3	PC^{3}, PC^{17}, PC^{20}, PC^{21}, PC^{23}, PC^{7}, PC^{4}, PC^{15}	71.80
S_4	PC^{8},PC^{3},PC^{17},PC^{2},PC^{19},PC^{4},PC^{10}	70.93
S_5	PC^{10}, PC^{17}, PC^{3}, PC^{2}, PC^{19}, PC^{4}, PC^{18}, PC^{11}, PC^{6}, PC^{23}, PC^{15}	71.76
S_{all}	All 23 physical–chemical properties	70.84

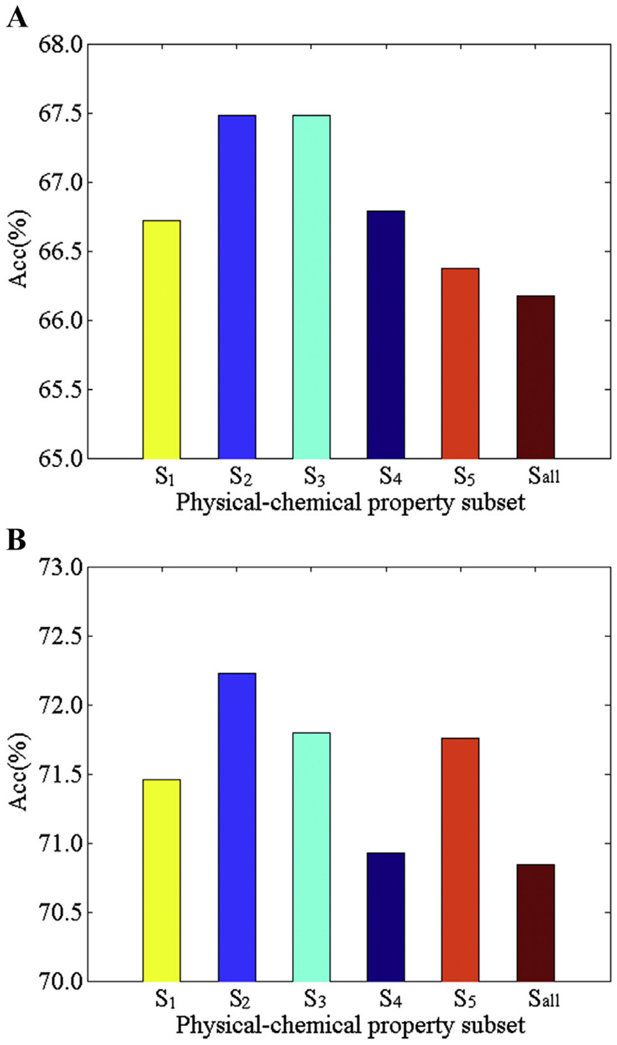

Fig.1. Comparisons of prediction accuracies (Acc) among the 5 generated candidate subsets and the full set of 23 physical–chemical properties: (A) under PseDNC feature representation; (B) under AC + CC feature representation.

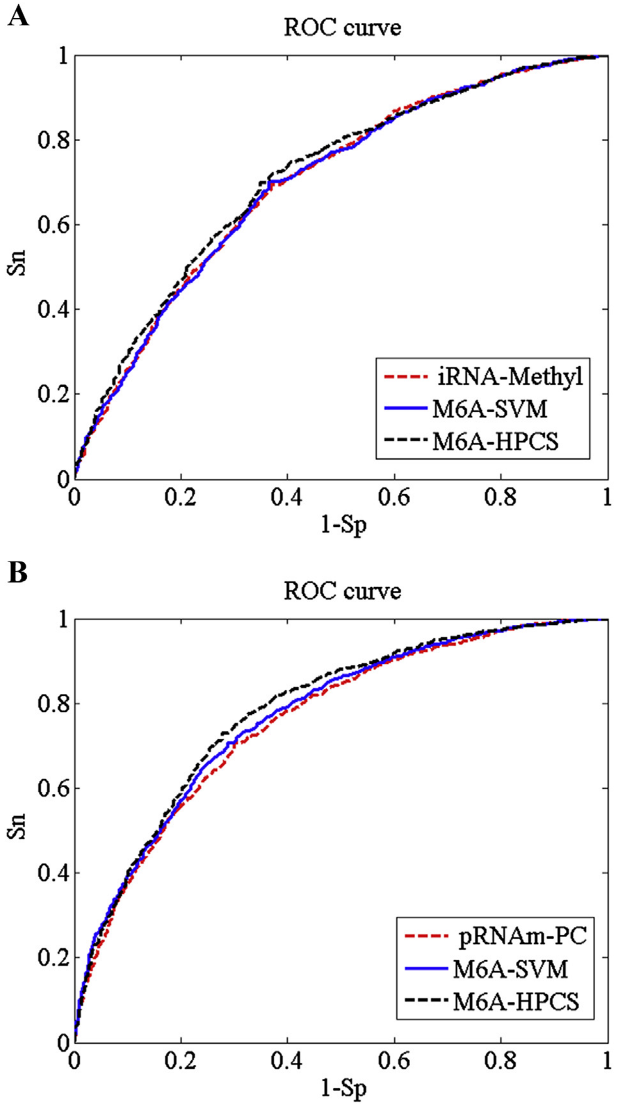

Fig.2. ROC curves of the different predictors under PseDNC (A) and AC + CC (B) feature representations.

Table 5
Comparisons among three m^6A site predictors—iRNA-Methyl, M6A-SVM, and M6A-HPCS—under PseDNC feature representation.

Predictor	Sp (%)	Sn (%)	Acc (%)	MCC	AUC	Optimized parameters
iRNA-Methyl[a]	60.63	70.55	65.59	0.29	0.705	$C = 32$, $\gamma = 0.0078$
M6A-SVM[b]	64.50	67.94	66.22	0.32	0.699	$C = 512$, $\gamma = 0.00098$
M6A-HPCS[c]	62.89	71.77	67.33	0.35	0.713	$C = 8$, $\gamma = 0.0625$

[a] Results excerpted from Ref. [28].
[b] Results obtained with the entire set of 23 physical–chemical properties.
[c] Results obtained with the optimized subset of the 23 physical–chemical properties.

In iRNA-Methyl [28], Chen and coworkers experimentally demonstrated that $\lambda = 6$ and $w = 0.9$ are better choices for performing m^6A site prediction. For the purpose of fair comparison, we took the same parameter setting of λ and w to implement M6A-SVM and M6A-HPCS.

From Table 5, it was found that the proposed M6A-HPCS outperforms iRNA-Methyl and M6A-SVM regarding the three overall evaluation indexes—Acc, MCC, and AUC—and acts as the best performer. The values of Acc, MCC, and AUC of M6A-HPCS are 67.33%, 0.35, and 0.713, respectively, which are 1.74, 6, and 0.8%

higher than those of iRNA-Methyl, which is a recently released m^6A site predictor. M6A-HPCS is also superior to M6A-SVM, with improvements of 1.11, 3, and 1.4% on Acc, MCC, and AUC, respectively, which further demonstrates the efficacy of the proposed physical–chemical property selection procedure under PseDNC feature representation.

2.3.2. Comparisons with predictors under AC + CC feature representation

We also performed comparisons among pRNAm-PC [27], M6A-SVM, and M6A-HPCS, all of which are under AC + CC feature representation and use SVM as the prediction engine. In pRNAm-PC [27], 10 physical–chemical properties were used to encode AC + CC features, whereas M6A-SVM and M6A-HPCS took the entire set and the optimized subset, respectively, of the 23 physical–chemical properties to encode RNA samples to AC + CC features. As to the parameter G in the AC + CC encoding scheme, we set its value to be 4 because pRNAm-PC also takes this parameter configuration. Table 6 summarizes the comparison results among pRNAm-PC, M6A-SVM, and M6A-HPCS under AC + CC feature representation. Fig. 2B plots the ROC curves of the three predictors under AC + CC feature representation.

From Table 6, it was found that the proposed M6A-HPCS acts as the best performer with Acc = 72.38%, MCC = 0.45, and AUC = 0.782, which are higher than those of pRNAm-PC. Again, M6A-HPCS outperforms M6A-SVM, indicating the efficacy of physical–chemical property selection under AC + CC feature representation.

In summary, the results listed in Tables 5 and 6 demonstrate that the proposed physical–chemical property selection algorithm is helpful for improving the performance of m^6A site prediction under both PseDNC and AC + CC feature representations. In addition, the implemented predictor M6A-HPCS, which is based on the proposed physical–chemical property selection algorithm, outperforms the existing state-of-the-art m^6A site predictors, including iRNA-Methyl [28] and pRNAm-PC [27].

3. Conclusion

Predicting m^6A sites fast and accurately solely from primary RNA sequences is useful for both basic research and drug development. Several recent studies [27,28] have revealed the feasibility of performing m^6A site prediction with physical–chemical properties of nucleotides but have not explained which physical–chemical properties are better choices. Inspired by these pioneering works, in this study we proposed a heuristic physical–chemical property selection algorithm that can optimize a subset from nucleotide physical–chemical properties under the prescribed feature representation to improve the performance of m^6A site prediction. Based on the proposed HPCS algorithm, we implemented a predictor, called M6A-HPCS, which can predict m^6A sites from RNA sequences with high accuracy. Experimental results on benchmark datasets have demonstrated that the proposed M6A-HPCS is superior to existing sequence-based m^6A site predictors, including iRNA-Methyl and pRNAm-PC. For the convenience of bioinformatics researchers, a web-server based on the HPCS algorithm with AC and CC feature representation has been put online and is available at http://csbio.njust.edu.cn/bioinf/M6A-HPCS. We believe that the proposed method will complement existing m^6A site predictors and benefit m^6A-related research studies.

Table 6
Comparisons among three m^6A site predictors—pRNAm-PC, M6A-SVM, and M6A-HPCS—under AC + CC feature representation.

Predictor	Sp (%)	Sn (%)	Acc (%)	MCC	AUC	Optimized parameters
pRNAm-PC[a]	69.75	70.55	69.74	0.40	0.763	$C = 32, \gamma = 0.0078$
M6A-SVM[b]	69.40	72.15	70.77	0.42	0.771	$C = 64, \gamma = 0.00098$
M6A-HPCS[c]	67.41	77.35	72.38	0.45	0.782	$C = 128, \gamma = 0.00098$

[a] Results excerpted from Ref. [27].
[b] Results obtained with the entire set of 23 physical–chemical properties.
[c] Results obtained with the optimized subset of the 23 physical–chemical properties.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61373062, 61233011, 61572242, and 61222306), the Natural Science Foundation of Jiangsu (BK20141403), the China Postdoctoral Science Foundation (2014T70526), the Fundamental Research Funds for the Central Universities (30916011327), the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), and "The Six Top Talents" of Jiangsu Province (2013-XXRJ-022).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ab.2016.06.001.

References

[1] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, Topology of the human and mouse m^6A RNA methylomes revealed by m6A-seq, Nature 485 (2012) 201–206.
[2] K.D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C.E. Mason, S.R. Jaffrey, Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons, Cell 149 (2012) 1635–1646.
[3] R. Desrosiers, K. Friderici, F. Rottman, Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells, Proc. Natl. Acad. Sci. U. S. A. 71 (1974) 3971–3975.
[4] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.-G. Yang, N^6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO, Nat. Chem. Biol. 7 (2011) 885–887.
[5] K. Karikó, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA, Immunity 23 (2005) 165–175.
[6] T.W. Nilsen, Internal mRNA methylation finally finds functions, Science 343 (2014) 1207–1208.
[7] Y. Niu, X. Zhao, Y.S. Wu, M.M. Li, X.J. Wang, Y.G. Yang, N^6-Methyl-adenosine (m^6A) in RNA: an old modification with a novel epigenetic function, Genomics Proteomics Bioinform 11 (2013) 8–17.
[8] Y. Yue, J. Liu, C. He, RNA N^6-methyladenosine methylation in post-transcriptional gene expression regulation, Genes Dev. 29 (2015) 1343–1355.
[9] Y. Saletore, K. Meyer, J. Korlach, I.D. Vilfan, S. Jaffrey, C.E. Mason, The birth of the Epitranscriptome: deciphering the function of RNA modifications,, Genome Biol. 13 (2012), http://dx.doi.org/10.1186/gb-2012-13-10-175.
[10] T. Chen, Y.-J. Hao, Y. Zhang, M.-M. Li, M. Wang, W. Han, Y. Wu, Y. Lv, J. Hao, L. Wang, m^6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency, Cell Stem Cell 16 (2015) 289–301.
[11] Y. Fu, D. Dominissini, G. Rechavi, C. He, Gene expression regulation mediated through reversible m^6A RNA methylation, Nat. Rev. Genet. 15 (2014) 293–306.
[12] S.D. Agarwala, H.G. Blitzblau, A. Hochwagen, G.R. Fink, RNA methylation by the MIS complex regulates a cell fate decision in yeast, PLoS Genet. 8 (6) (2012) e1002732.
[13] G. Jia, Y. Fu, C. He, Reversible RNA adenosine methylation in biological regulation, Trends Genet. 29 (2013) 108–115.
[14] J.-M. Fustin, M. Doi, Y. Yamaguchi, H. Hida, S. Nishimura, M. Yoshida, T. Isagawa, M.S. Morioka, H. Kakeya, I. Manabe, RNA-methylation-dependent RNA processing controls the speed of the circadian clock, Cell 155 (2013) 793–806.
[15] S. Schwartz, S.D. Agarwala, M.R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T.S. Mikkelsen, R. Satija, G. Ruvkun, S.A. Carr, E.S. Lander, G.R. Fink, A. Regev, High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis, Cell 155 (2013) 1409–1421.
[16] A. Klungland, J.A. Dahl, Dynamic RNA modifications in disease, Curr. Opin. Genet. Dev. 26 (2014) 47–52.

[17] S. Blanco, M. Frye, Role of RNA methyltransferases in tissue renewal and pathology, Curr. Opin. Cell Biol. 31 (2014) 1–7.

[18] G. Zheng, J.A. Dahl, Y. Niu, P. Fedorcsak, C.-M. Huang, C.J. Li, C.B. Vågbø, Y. Shi, W.-L. Wang, S.-H. Song, ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility, Mol. Cell 49 (2013) 18–29.

[19] H. Yang, Y. Zheng, T.W. Li, H. Peng, D. Fernandez-Ramos, M.L. Martínez-Chantar, A.L. Rojas, J.M. Mato, S.C. Lu, Methionine adenosyltransferase 2B, HuR, and sirtuin 1 protein cross-talk impacts on the effect of resveratrol on apoptosis and growth in liver cancer cells, J. Biol. Chem. 288 (2013) 23161–23170.

[20] Z. Zhu, B. Wang, J. Bi, C. Zhang, Y. Guo, H. Chu, X. Liang, C. Zhong, J. Wang, Cytoplasmic HuR expression correlates with P-gp, HER-2 positivity, and poor outcome in breast cancer, Tumor Biol. 34 (2013) 2299–2308.

[21] J. Meng, Z. Lu, H. Liu, L. Zhang, S. Zhang, Y. Chen, M.K. Rao, Y. Huang, A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package,, Methods 69 (2014) 274–281.

[22] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of N^6-methyladenosine by m^6A-seq based on immunocapturing and massively parallel sequencing, Nat. Protoc. 8 (2013) 176–189.

[23] Y. Li, X. Wang, C. Li, S. Hu, J. Yu, S. Song, Transcriptome-wide N^6-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification, RNA Biol. 11 (2014) 1180–1188.

[24] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, J. Biomol. Struct. Dyn. 33 (2015) 1731–1742.

[25] X. Xiao, M.-J. Hui, Z. Liu, W.-R. Qiu, iCataly-PseAAC: identification of enzymes catalytic sites using sequence evolution information with grey model GM (2,1), J. Membr. Biol. 248 (2015) 1033–1041.

[26] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.

[27] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, K.-C. Chou, pRNAm-PC: predicting N^6-methyladenosine sites in RNA sequences via physical–chemical properties, Anal. Biochem. 497 (2015) 60–67.

[28] W. Chen, P. Feng, H. Ding, H. Lin, K.-C. Chou, iRNA-Methyl: identifying N^6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.

[29] Z. Liu, X. Xiao, W.-R. Qiu, K.-C. Chou, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.

[30] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, Biomed. Res. Int. 2014 (2014) 947416.

[31] W. Chen, H. Tran, Z. Liang, H. Lin, L. Zhang, Identification and analysis of the N^6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome, Sci. Rep. 5 (2015) 13859.

[32] A. Perez, A. Noy, F. Lankas, F.J. Luque, M. Orozco, The relative flexibility of B-DNA and A-RNA duplexes: database analysis,, Nucleic Acids Res. 32 (2004) 6144–6151.

[33] J.R. Goñi, A. Pérez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol. 8 (2007) R263.

[34] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, D.H. Turner, Improved free-energy parameters for predictions of RNA duplex stability, Proc. Natl. Acad. Sci. U. S. A. 83 (1986) 9373–9377.

[35] P. Ponnuswamy, M.M. Gromiha, On the conformational stability of oligonucleotide duplexes and tRNA molecules, J. Theor. Biol. 169 (1994) 419–432.

[36] M. Friedel, S. Nikolajewa, J. Sühnel, T. Wilhelm, DiProDB: a database for dinucleotide properties,, Nucleic Acids Res. 37 (2009) D37–D40.

[37] I. Barzilay, J. Sussman, Y. Lapidot, Further studies on the chromatographic behaviour of dinucleoside monophosphates, J. Chromatogr. A 79 (1973) 139–146.

[38] M.M. Gromiha, Development of RNA stiffness parameters and analysis on protein–RNA binding specificity: comparison with DNA, Curr. Bioinform 7 (2012) 173–179.

[39] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[40] K.-C. Chou, H.-B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (2007) 633–640.

[41] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.

[42] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.

[43] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

[44] 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.

[45] P. Feng, W. Chen, H. Lin, Prediction of CpG island methylation status by integrating DNA physicochemical properties, Genomics 104 (2014) 229–233.

[46] M. Kabir, M. Iqbal, S. Ahmad, M. Hayat, iTIS-PseKNC: identification of translation initiation site in human genes using pseudo *k*-tuple nucleotides composition, Comput. Biol. Med. 66 (2015) 252–257.

[47] 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, Anal. Biochem. 456 (2014) 53–60.

[48] 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, Anal. Biochem. 462 (2014) 76–83.

[49] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.

[50] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn 20 (1995) 273–297.

[51] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, F. Tahi, Towards a piRNA prediction using multiple kernel fusion and support vector machine, Bioinformatics 30 (2014) i364–i370.

[52] D.-J. Yu, J. Hu, J. Yang, H.-B. Shen, J. Tang, J.-Y. Yang, Designing template-free predictor for targeting protein–ligand binding sites with classifier ensemble and spatial clustering, IEEE/ACM Trans. Comput. Biol. Bioinform 10 (2013) 994–1008.

[53] 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, J. Comput. Chem. 34 (2013) 974–985.

[54] K.-C. Chou, Y.-D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.

[55] Y.-D. Cai, G.-P. Zhou, K.-C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.

[56] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, New York, 2000.

[57] C.C. Chang, C.J. Lin, LIBSVM: a Library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2006) 389–396.

[58] R.E. Fan, P.H. Chen, C.J. Lin, Working set selection using second order information for training SVM, J. Mach. Learn. Res. 6 (2005) 1889–1918.

[59] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.

[60] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[61] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo, K.-C. Chou, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (10) (2012) e47843.

[62] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68.

[63] W.-R. Qiu, X. Xiao, K.-C. Chou, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.

[64] M. Kabir, M. Hayat, iRSpot-GAEnsC: identifying recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, Mol. Genet. Genomics 291 (2016) 285–296.