Abstract

The N6-methyladenosine (m6A) is a type of post-transcription modifications that occurs on the sixth position of nitrogen atoms of adenosine. It is the most widely present modification among all RNA modifications. The m6A modification was discovered in 1970s and can be found approximately in all types of cellular RNAs, such as, ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA) and long non-coding RNA (ncRNA). The m6A modifications control various biological processes, such as, gene regulation, micro-RNA regulation, X-chromosome inactivation, cell reprogramming, cell differentiation and RNA localization, etc. Any abnormal mutation in the m6A site may lead to a number of conditions, including, brain-related disorders, prostate cancer, stomach cancer, kidney cancer, pancreatic cancer, Leukemia, sarcoma, mesothelioma and many other diseases. Therefore, correct identification of m6A sites is essential to overcome these diseases. Additionally, different RNA viruses, such as, herpes virus, simian virus 40 (SV40), influenza virus, adenovirus and Rous sarcoma virus etc., also contain internal m6A sites in their RNA transcriptomes. More recently, researchers have been working on the correct identification of m6A sites, especially in the Saccharomyces cerevisiae species. However, existing methods are facing numerous challenges in the prediction of m6A sites in Saccharomyces cerevisiae species. First of all, there are complex sequence patterns surrounding the m6A sites, which are not dealt in detail by the existing methods. The second reason for downgraded performance is the use of mainly statistical or chemical properties based features of nucleotides, while there are other interesting aspects of nucleotides which remain unexplored. Likewise, the third problem relates to m6A site patterns hidden in the long-range local sequence order which are not well grounded in the existing models. In this work, we seek to address all these problems by designing and evaluating three predictors namely, 1) m6A-Pred, 2) m6A-Finder and 3) m6A-Class. The m6A-Pred method is based on the fusion of features crafted using statistical and chemical properties of nucleotides. The feature fusion improved the m6A site detection by efficiently capturing complex patterns surrounding the m6A sites. The

m6A-Pred detects potential sites using the Random forest classifier and it is benchmarked on a standardized dataset containing a total of 2,614 samples. The m6A-Pred outperformed existing techniques and achieved a sensitivity of 79.65%, an accuracy of 78.58% and a Matthew correlation coefficient (MCC) of 0.5717. Further improvements for the detection of m6A sites were done by designing a second predictor called the m6A-Finder. The m6A-Finder integrates uniquely constructed local as well as global sequence order features. This method uses hexa-nucleotide composition features as a long-order local sequence order information and auto-correlation features based on physical properties as a global order information. The Support vector machine (SVM) classifier is used to predict whether the input sequence has m6A site or not. The m6A-Finder outperforms all of the existing predictors by achieving a sensitivity of 82.10%, a specificity of 81.94%, an accuracy of 82.02%, and a MCC score of 0.6404. Finally, we address all stated problems in our last method called the m6A-Class predictor. The m6A-Class is based on the fusion of local and global sequence order. For local sequence order, the hepta-nucleotide composition features, as well as chemical property based features are used, while for global sequence order auto-covariance features based on physical properties are used. The m6A-Class is more accurate in the prediction of m6A sites because it captures details of a sequence in all aspects i.e statistical, chemical and physical. For site detection, the SVM classifier is employed to predict the presence or absence of an m6A site. The m6A-Class method outperforms all the existing predictors by achieving a sensitivity of 88.45%, a specificity of 86.30%, an accuracy of 87.38%, and an MCC score of 0.7477. Lastly, from this work we conclude that the proposed predictors are well-suited for the prediction of m6A sites in RNA transcripts especially for the detection of m6A sites in the Saccharomyces cerevisiae species.

Contents

List of Publications

1. **A. Khan**, H. U. Rehman, U. Habib, and U. Ijaz, "Detecting n6-methyladenosine sites from rna transcriptomes using random forest," Journal of Computational Science, vol. 47, p. 101238, 2020. **(Impact Factor: 2.64)**
2. **A. Khan**, H. U. Rehman, U. Habib, and U. Ijaz,"m6A-Finder: Detecting m6A methylation sites from RNA transcriptomes using physical and statistical properties based features,Computational Biology and Chemistry vol. , p. , 2021.(under-review) **(Impact Factor: 2.90)**
3. **A. Khan**, H. U. Rehman, U. Habib, and U. Ijaz,"m6A-Class: Identification of m6A sites in Saccharomyces cerevisiae using reduced hybrid features," Journal of Computational Science, vol. , p. , 2021.(under-review) **(Impact Factor: 2.64)**

List of Figures

List of Tables

ACADEMIC SOLUTIONS

Chapter 1

Introduction

In the central dogma of molecular biology, the expression of a gene results in both coding and non-coding RNA molecules. The full range of all coding RNA molecules in a single cell or multiple cells is known as transcriptome. A transcriptome comprises of all RNA transcripts in a particular species which make proteins. The process of finding genetic codes in transcriptomes is called transcriptome sequencing. Different factors like external environments and internal mutations can alter transcriptome sequences. Proteins, DNA and RNA all have their specific modification sites. Although proteins and DNA modifications are extensively studied by researchers, but they have limited understanding of messenger RNA modifications because the field is new and supporting evidences are evolving with the increase in available data.

Among all discovered epigenetic modifications of cellular RNA, the N6-methyladenosine (m6A) is the most frequently occurring modification [1]. The m6A modification occurs on sixth position of nitrogen atoms of adenosine nucleotide. It is discovered in the 1970s and can be found in most eukaryotes, such as, plants, insects, mammals and yeasts. It can be found approximately in all type of RNAs, such as, tRNA, rRNA, snRNA and long non-coding RNA (lncRNA). The N6-adenosyl methyltransferase complexes including METTL3, METTL14 and WTAP [2], catalyze these modifications because they have specific binding sites for these methyltransferases.

The m6A modifications govern different biological processes, such as, gene regulation, micro-RNA regulation, X-chromosome inactivation, cell reprogramming, cell differentiation, protein localization and protein translation [3]. These biological processes are

mediated by a specific type of RNA binding proteins which specifically recognize these sites. These m6A binding proteins are known as m6A readers. The RNA binding proteins, such as, YTHDF1, YTHDF2, YTHDF3 and YTHDC1 are characterized as m6A readers. Any abnormal mutation in the m6A site make it difficult for readers to be attached. This situation may lead to brain-related disorders, breast cancer [4], leukemia [5], prostate cancer [6], thyroid tumor [7] pancreatic cancer, sarcoma, and many other diseases [8]. Moreover, a recent study also claims that m6A sites are non-randomly distributed across genomes [9], making it difficult to predict m6A sites. Therefore, to overcome the above stated problems, the correct identification of m6A sites has become very crucial.

1.1 Background

Recently, web lab experimental techniques were utilized to find a single-nucleotide resolution map of the m6A sites across human transcriptome. Their results are reasonable for human, but for other species, they have a high false-positive and false-negative rate [10]. On the other hand, wet lab experimental techniques are time consuming and expensive. Therefore, we need computational techniques to accurately distinguish m6A site from non-m6A sites in the transcriptome. Apart from finding m6A sites correctly, these techniques will save time and cost.

Nowadays, with the help of high-throughput techniques, complete transcriptomic sequences are available for human, plants, yeast and other mammals etc. [9, 11, 12]. Many machine learning-based computational tools are developed based on this data. Although, these tools performed well on mammalian data, when they evaluated the Saccharomyces cerevisiae transcriptomic sequences, they have a high miss prediction rate. The main reason behind this low performance is that the hidden useful patterns surrounding m6A sites in Saccharomyces cerevisiae are not explored properly.

Many researchers utilized statistical as well as some of the chemical properties based features to distinguish m6A sites from non-m6A sites [13, 14], but these features are not sufficient to differentiate m6A sites from non-m6A sites in Saccharomyces cerevisiae transcriptome due some problems: (1) There are complex patterns surrounding m6A sites.

(2) The second reason for high false positive/negative is that they utilize chemical or statistical features. (3) The third reason is that, they use short-range local sequence order features. They neither utilize long-range local sequence order features independently nor with combination of global features.

To overcome the above problems, in this work, we propose three predictors namely m6A-Pred, m6A-Finder and m6A-Class. The m6A-Pred solves problem 1 and problem 2 by using fusion of statistical and chemical properties based features. Although a fusion of features captures hidden useful insights surrounding m6A sites, but also increases vector dimension, which causes overfitting. To solve this problem, we use Genetic Algorithm (GA) to select optimal features. Finally, we use Random forest for classification. The m6A-Finder and m6A-Class address problem 3 by using long-range local sequence order features (hexa-nucleotide composition) with global features (autocorrelation features based on physical properties). Although, long-range local sequence order features capture useful local details of sequence, but also increase vector dimension. To solve this problem, we use Minimum redundancy maximum relevance (mRMR) to select relevant features and discard redundant ones. Finally, we use Support vector machine for prediction. The m6A-Class uses three types of features i.e statistical, chemical and physical to capture hidden information surrounding m6A sites more accurately. As we know that fusion of these diverse features increases vector dimension which leads to overfitting. To overcome this issue, we use mRMR to select relevant features and discard redundant ones. Finlly, we use SVM to predict the absence or present of m6A site in a input sequence. Our predictors outperform state-of-the-art predictor in performance. The m6A-Pred outperforms in sensitivity, accuracy and MCC from all current state-of-the-art predictors. The m6A-Finder and m6A-Class outperform in all four metrics. Finally, it is concluded that the developed predictors can be used as a useful biomedical tool for the prediction of m6A sites in various species; thus benefiting in basic research and drug discovery for various diseases.

1.2 Problem Statement

The existing methods face various challenges in the identification of m6A sites in saccharomyces cerevisiae species, which are:

1. There are complex sequence patterns surrounding the m6A sites, which are not dealt in detail by the existing methods.
2. Existing predictors either utilize statistical features or chemical features to detect N6-methyleadenosine (m6A) sites in saccharomyces cerevisiae species which results in false predictions. The fusion of these two types of features has not been widely explored. In addition to this, there are other interesting aspects of nucleotides which remain unexplored.
3. The third problem relates to m6A site patterns hidden in the long-range local sequence order which are not well grounded in the existing models.

1.3 Research Questions

1. Can physical properties based features enhance the performance of the m6A site prediction algorithms?
2. Can physical properties based feature with combination of long range local sequence order features enhance the performance of the m6A site predictor?

1.4 Contribution

In this work, we seek to address the problems highlighted in section 1.2 by designing and evaluating three predictors namely, 1) m6A-Pred, 2) m6A-Finder and 3) m6A-Class. The overall research contributions of this thesis are summarized as follows:

1. In the first method, we extract and combine features based on statistical and chemical aspects of nucleotides from input RNA sequences. The Genetic algorithm is

utilized for the first time for this problem to remove unnecessary features and to select optimal features.

2. In the second method, we design features based on hexa-nucleotide composition along with physical properties based global features. The Minimum redundancy maximum relevance (mRMR) is used to select relevant features and remove redundant and un-necessary features.

3. In the third method, we construct features based on hepta-nucleotide composition, chemical properties based features and physical properties based global features. The Minimum redundancy maximum relevance (mRMR) is used to select relevant features and remove redundant and un-necessary features.

1.5 Thesis organization

The remaining part of this thesis is organized as follows: Chapter 2 provides a comprehensive literature review about current state-of-the-art techniques for m6A sites prediction in yeast RNA sequences. A proposed model with the title, "Detecting N6-methyladenosine sites from RNA transcriptomes using Random Forest" is presented in Chapter 3. In chapter 4, a detailed explanation is given for the second predictor we developed, with the title "m6A-Finder: Detecting m6A methylation sites from RNA transcriptomes using physical and statistical properties based features". We describe the third predictor in Chapter 5 with the title: "m6A-Class: Identification of m6A sites in Saccharomyces cerevisiae using reduced hybrid features". Finally, in Chapter 6, we conclude the thesis and discuss some future directions for the current work.

ACADEMIC SOLUTIONS

Chapter 2

Related Work

N6-methyladenosine (m6A) is one of the abundant post-transcription modification in cellular RNA. It regulates different biological processes, such as, protein synthesis, X-chromosome inactivation, cell stability, cell-reprogramming and miRNA regulation etc. Most recently, various studies claimed that mutations in m6A sites are linked with various diseases, such as, brain-tumor, heart attack, obesity and cancer. The correct identification of m6A sites is essential to overcome these diseases. However, the state-of-the-art predictors face many challenges for precise detection of m6A sites. Even for model organisms, such as Saccharomyces cerevisiae, the detection of m6A sites is difficult due to complex patterns surrounding the m6A sites. These patterns are not widely understood and lead to non-discriminative features for detecting m6A sites. To overcome this problem, novel computational tools are needed to detect m6A sites more accurately.

2.1 Wet-lab based m6A Prediction

Many researchers utilized the wet-lab techniques to detect m6A sites in the human transcriptome. They used a single-nucleotide -resolution map for showing these sites across the genome. This technique showed better results for human transcriptome. When the mentioned method was tested on other transcriptomes, it failed to detect m6A sites. Besides that, they used "high-performance liquid chromatography" technique [15], "thin

layer two-dimensional chromatography" technique [16], next-generation sequencing technique [17] to find m6A sites in various transcriptomes. All of them suffered from high false-positive, and false-negative rate in other species [10]. In addition to high false positive and false negative detections, these techniques are expensive and laborious. Therefore, we require computational techniques to save time as well as cost.

2.2 Computational Approaches for m6A Prediction

Due to current advancement in sequencing technologies, the genome-wide distributions for m6A sites are available for various species, such as, human, A. thaliana, yeast and, some other species [9, 11, 12]. This large amount of data helped researchers make computational predictors to separate m6A sites from non-m6A sites. A lot of computational tools based on machine learning were proposed for the detection of m6A sites in different species,such as,yeast,human, mouse,etc.

In [13], Chen and his co-authors proposed a predictor named iRNA-Methyl to distinguish m6A sites from non-m6A sites in Saccharomyces cerevisiae species. The dataset used for experimental purpose contains a total of 2614 sequences. Half of them have m6A sites, while half of them have non-m6A sites. In [13], they extracted features based on Pseudo nucleotide composition technique (PseNC). Finally, they utilized Support vector machine(SVM) for classification. The iRNA-Methyl has good sensitivity but low specificity due to simple statistical features, which are not enough discriminate between non-m6A sites and m6A sites.

In m6Apred [18], researchers used chemical properties based values to extract features from nucleotide. Finally, Support vector machine was selected as a classifier. The m6Apred has good specificity but low sensitivity on an independent dataset. The reason behind low sensitivity is that chemical features captures only short order local details which are not enough for differentiation between m6A sites and non-m6A sites.

Motivated by [18], a group of researchers proposed a predictor named pRNAm-PC [19]. They injected chemical properties based values of nucleotides into auto-covariance and cross-variance to extract features from a given sequence. The proposed predictor has two

downsides: it has a high dimensional feature vector and low accuracy.

Another predictor named SRAMP [20] was developed for the detection of m6A sites in multiple species data. The authors of the paper used position-specific dinucleotide encoding scheme and secondary structure patterns for feature extraction. The proposed predictor achieved low sensitivity compared to previously developed predictors. Besides, the performance of the predictor is not good on yeast data. The yeast species has complex patterns in structure and position-specific dinucleotide encoding scheme is not sufficient to capture these patterns.

To make further improvement in the classification of post-transcription modification sites, another predictor named MethyRNA [21] was proposed. The authors in this paper extracted chemical property based features. The proposed predictor achieved better results in detection m6A sites in mammalian dataset but when it was tested on yeast data the results were not so satisfactory [22] . The mentioned feature extraction techniques are not sufficient to clearly distinguish m6A sites from non-m6A sites in yeast species because of its complex structure from other spices.

A group of researcher used a heuristic approach to select subset chemical features from multiple sets of chemical property based features [23]. The proposed predictor has good performance in term of accuracy compared to other state-of-the-art predictors, but its specificity was low. The main reason of low specificity is that they used only chemical properties based features, which can not accurately separate non-m6A sites from m6A sites.

More recently, another group of researchers proposed another predictor called RAM-ESVM [24] for the classification m6A sites and non-m6A sites in yeast transcriptome. Researchers in this approach used pseudo-di-nucleotide composition technique, motif technique and gapped K-mer technique for feature extraction. They trained three independent support vector machines on these three types of features. Finally, majority voting was used for the final prediction. The performance of the proposed predictor is acceptable compared to existing predictors. They used only local order statistical features which are not enough for accurate classification of m6A sites in yeast. They did utilized chemical properties based features with statistical features.

Inspired by [20] work, a predictor named imethyl-STTNC [25] was developed to predict m6A sites in yeast and human data. The authors in the mentioned paper used four types of feature encoding schemes, which are Pseudo Di-nucleotide Composition (PseDNC), Pseudo Tri-nucleotide composition (PseTNC), split Tri-nucleotide composition and split Tetra-nucleotide composition. Finally, they used the support vector machine for classification purpose. The proposed predictor performed well on the human dataset, but its results on the yeast dataset are not satisfactory. They used only local short-order statistical features, which are not sufficient to differentiate between m6A sites and non-m6A sites in yeast.

Similar to [20], BERMP [26] was developed to predict m6A sites in multiple species data. In the proposed predictor, the authors used Enhanced Nucleic Acid Composition (ENAC) as a feature extraction technique. Finally, the prediction was made via an ensemble classification of random forest and deep learning. The predictor achieved good accuracy on mammalian and plant data. The proposed predictor did not perform well on yeast data. The features they used are short order local features, which are not enough to capture complex patterns surrounding m6A sites in yeast.

Like BERMP [26], a predictor named M6AMRFS [14] was proposed for the detection of m6A sites in yeast, human, mouse and A. Thaliana. In the mention predictor, researchers used Di-nucleotide composition features based on local density information and binary encoding of nucleotides as feature representation schemes. The Extreme gradient boost was used as a classifier. The predictor performed well on human, mouse and A. Thaliana. The M6AMRFS has average performance on yeast data and still has a gap for further improvement. They used only local short-order sequential features which are not sufficient for accurate classification of m6A sites.

Recently, a predictor named DeepM6ASeq [27] for the identification of m6A sites in mammalian species data. A given input sequence was converted into a numerical vector using a one-hot encoding scheme. Finally, A Convolutional Neural Network (CNN) is utilized for classification. The proposed predictor performed well on mammalian data and achieved good accuracy, sensitivity, specificity. When DeepM6ASeq is evaluated on yeast data, the sensitivity was low compared to existing state-of-the-art predictors[28].

The encoding scheme they used was not enough to distinguish between m6A sites and non-m6A sites.

Further improvement for m6A sites prediction, were proposed by a few researchers with the name called SICM6A [28]. They extracted features using 3-mer techniques from given RNA sequence. Finally, Gated Recurrent Unit (GRU) based deep learning approach was used for classification. The performance of the proposed predictor was good on plant and mammalian data. When it was tested on Saccharomyces cerevisiae data its performance was approximately equal to BERMP [26]. The mentioned feature extraction scheme is not enough to capture complex hidden patterns in yeast species.

Similar to pRNAm-PC [19], a predictor named iRNA-Freq [29] was proposed for prediction m6A sites in Saccharomyces cerevisiae data. The frequent gap-Kmer approach was used as a feature extraction scheme. Finally, they used a linear regression-based classifier to predict whether the given sequences contain m6A site or not. The reported results show that iRNA-Freq has low specificity compared to existing state-of-the-art techniques. The gap-Kmer technique captures only local statistical details which are not sufficient to discriminate non-m6A sites from m6A sites.

Most recently, iMRM [30] was developed to find m6A sites in five types of RNA modification sites in human, mouse and yeast data. They extracted features from a given sequence using local sequence order feature techniques. The results of the mentioned predictor are acceptable for mouse and human modification sites. It has low accuracy in finding m6A sites in yeast transcriptome compared to other existing predictors.

From a detailed review, it may be concluded that feature extraction methods have a vital role in the performance of a predictor. However, existing methods are facing numerous challenges in the prediction of m6A sites in Saccharomyces cerevisiae species. First of all, there are complex sequence patterns surrounding the m6A sites, which are not dealt in detail by the existing methods. The second reason for downgraded performance is the use of mainly statistical or chemical properties based features of nucleotides, while there are other interesting aspects of nucleotides which remain unexplored. Likewise, the third problem relates to m6A site patterns hidden in the long-range local sequence order which are not well grounded in the existing models. In this work, we seek to address

all these problems by designing and evaluating three predictors namely, 1) m6A-Pred, 2) m6A-Finder and 3) m6A-Class. The conceptual novel of this work lies in using novel nucleotide properties based features along with fusion of various types of features for the prediction of m6A sites in Saccharomyces cerevisiae species.

Chapter 3

Detecting N6-methyladenosine sites from RNA transcriptomes using Random Forest

3.1 Introduction

Among the 150 types of post transcription modifications of cellular RNA, the N6-methyladenosine (denoted as m6A) is the most frequently occurring modification [31]. These modifications are catalyzed by N6-adenosyl methyltransferase complexes including METTL3, METTL14 and WTAP and some others [1, 2]. The m6A sites are present in both the prokaryotes and eukaryotes [1]. In a recent study, it is found that m6A modifications control different types of biological processes like localization and translation of proteins and some other essential cellular tasks [3]. Any abnormal change in m6A may lead to certain abnormalities, including, cancer, brain related disorders , and a lot of other diseases [8, 32]. Furthermore, it is also found that m6A is non-randomly distributed across genomes [9]. Thus, identification of m6A sites is important in order to understand many key biological functions.

Traditionally, wet lab experiments are used for providing single-nucleotide resolution map of the m6A sites across human transcriptomes. However, the results of these resolution maps for other species m6A sites are not satisfactory because of false positive and false

negative detections [10]. In addition, genome-wide m6A site detection through wet experiments is laborious and costly. Thus, we require computational tools to precisely detect the m6A sites. These computational tools will find m6A sites as well as save experimental time and costs. The benefits of these tools will be a direction for future research in bioinformatics and drug discovery for severe diseases like cancer, brain disorder and other abnormalities. Above aforementioned benefits motivated us to make a predictor for classification of m6A sites.

Due to efficient high-throughput technologies, the m6A genome-wide distributions are existing for different types of species like Homo sapiens, Arabidopsis thaliana, Saccharomyces cerevisiae and likewise other species [9, 11, 12]. Researchers got unprecedented opportunities due to the availability of large volume experimental data. These type of data provide the feasibility for developing computational tools in order to correctly predict m6A sites. Different type of computational methods were proposed according to the nature of data for the identification of m6A sites. For example, in [13, 18] authors proposed two computational tools for yeast specific data. In the first method, the authors used nucleotide chemical property to encode features, while in the second method the authors use pseudo nucleotide composition (PseNC) for encoding the RNA sequence. The first one has low sensitivity on independent dataset, while the second one has low specificity.

Another group of researchers, developed a predictor named pRNAm-PC by using chemical property based auto-covariance and auto cross-variance features in pseDNC [19]. The proposed technique has two problems, first, the feature vector dimension is very high and secondly, the accuracy they reported is considerably low. Another predictor tool namely, SRAMP was proposed for mammalian m6A sites classification [20]. Their work was inspired from the works of [13, 18] which primarily uses binary encoding scheme for feature extraction. The specificity of the proposed work is good but sensitivity is very low.

Subsequently, some researchers in [21] developed another predictor MethyRNA for the prediction of m6A site in Homo sapiens and Mus musculus datasets. While the performances of these proposed techniques are good for finding m6A sites in mammalian transcriptomes, but their results are not good to identify m6A sites accurately in yeast

transcriptome [20, 21]. The use of these features result in inefficient detection of m6A sites in yeast transcriptomes due to different structure from other species and inability of the computational techniques to capture the encoded information surrounding the m6A sites. Moreover, the information around yeast m6A sites are not well studied [20].

Some authors introduced heuristic based nucleotide physical-chemical property selection algorithm for m6A sites identification. The results of this technique are good from previous techniques in term of sensitivity but specificity is still low [23]. Recently, further improvements were carry out by [24], using ensemble support vector machines. In their work, they used PseKNC (pseudo K-tuple Nucleotide Composition), along with motif and gapped k-mer based techniques for feature extraction. The final output is generated by using majority voting strategy. The accuracy is good but time complexity is high due to high dimensional feature vectors as well ensemble classifiers.

In [25], the authors introduced a new predictor named imethyl-STTNC. They used split trinucleotide composition and split tetranucleotide composition features for RNA sequence encoding. Multiple classifiers were used for the detection of m6A sites in RNA sequences. The support vector machine achieved high performance on these features. The reported accuracy is good for Homo sapiens dataset but not satisfactory for yeast dataset. Recently, in [26], a new predictor called BERMP has been introduced for classification of m6A sites of multiple species. The proposed predictor's accuracy is good for mammalian and plant species. In this work, the authors used ENAC (Enhanced Nucleic Acid Composition) as feature encoding scheme. They used deep learning and Random forest based ensemble classifier for prediction, while the prediction accuracy still did not improve for yeast species RNA sequences.

Most recently, M6AMRFS was introduced for the prediction m6A sites in multiple species namely, Saccharomyces cerevisiae, Homo sapiens, Musculas and Arabidopsis Thaliana, [14]. The authors used local position specific dinucleotide frequency composition and binary encoding to encode features. They used Extreme Gradient boost (XG boost) classifier. However, its performance in identification of m6A sites in yeast transcriptome remained average, and further enhancements are still needed.

A predictor named DeepM6ASeq was proposed in [27] for classification m6A sites. They

used CNN for the detection of m6A sites. But the predictor is only evaluated for mammalian data. In recent time, another predictor WHISTLE, [33] was developed. The authors extracted sequence derived features and genome derived features from human mature messenger RNA sequences and full transcript sequences. They did not test their predictor on yeast data. Mostly recently, a predictor SICM6A[28] was developed for cross species m6A data classification. However, the accuracy of the mentioned predictor is good for all species, but in yeast transcriptomes, its results are nearly equal to [26].

From the extensive literature review above, we conclude that, the feature representation techniques used, are not satisfactory for the identification of m6A sites in yeast species. Moreover, apart from statistical and chemical properties of nucleotides in transcriptomes, another aspect that is not widely explored is the fusion of these characteristics surrounding the m6A sites. In this work, we propose a novel method (called *m6A-pred* predictor) that utilizes a fusion of characteristics including, statistical, and chemical properties of the nucleotides, to precisely predict the presence of m6A sites in RNA sequences. The proposed method will overcome the problem of inefficient detection of m6A sites in yeast transcriptomes (due to varied structure) and inability of the computational techniques to capture the encoded information surrounding the m6A sites. The extracted hybrid features are usually high dimensional. To reduce feature dimensionality, we also explore the feature importance through feature selection methods.

The main contributions of this study are as follow:

1. Combination of chemical properties based features and statistical based features are extracted from RNA to capture sounding information of m6A sites in more efficient way.

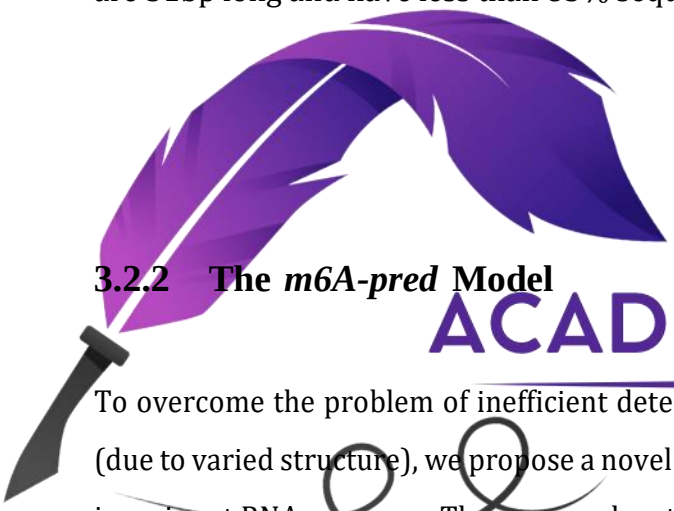
2. Wrapper based feature selection method is used to select important features and remove redundant and unnecessary features.

The remaining part of the paper is organized as follows: in Section 2, we detail the materials and method used to develop the proposed *m6A-pred* predictor. We benchmark our results in Section 3, on the intriguing RNA transcripts of *Saccharomyces cerevisiae* species. Lastly, we conclude the study in section 4.

3.2 Materials and Methods

3.2.1 Benchmark Dataset

The benchmark dataset to conduct this study is taken from [13]. The data set contains 2614 sequences, which has 1,307 positive sequences (having m6A sites) and 1,307 negative sequences (having non-m6A sites). The positive subset of the whole dataset are experimentally identified m6A sites. To prevent imbalance bias in training set, these 1,307 non-m6A samples were randomly selected from the 33,280 non-m6A sites. All sequences are 51bp long and have less than 85% sequence similarity.



3.2.2 The *m6A-pred* Model

To overcome the problem of inefficient detection of m6A sites in yeast transcriptomes (due to varied structure), we propose a novel method to predict the presence of m6A site in an input RNA sequence. The proposed method consists of three steps, namely, *feature extraction*, *feature selection*, and *classification*. In the first step, three different types of hybrid features will be extracted from the input RNA sequence. The novelty of this work lies in constructing features based on the fusion of chemical and statistical characteristics of the individual nucleotides to predict the presence of m6A sites. The crafted features are of different types, resulting in high dimensional vectors. These vectors are combined into a single feature vector. In the second step, this high dimensional vector will be processed using feature selection techniques and dimensions will be reduced by pruning the less significant dimensions. Finally, in step three, the resultant low dimension feature vector will be given as input to a random forest classifier, to predict weather the input sequence has an m6A sites or not. The model of the proposed predictor is graphically represented in Figure 3.1. The steps involved in the *m6A-pred* predictor are briefly elaborated bellow:

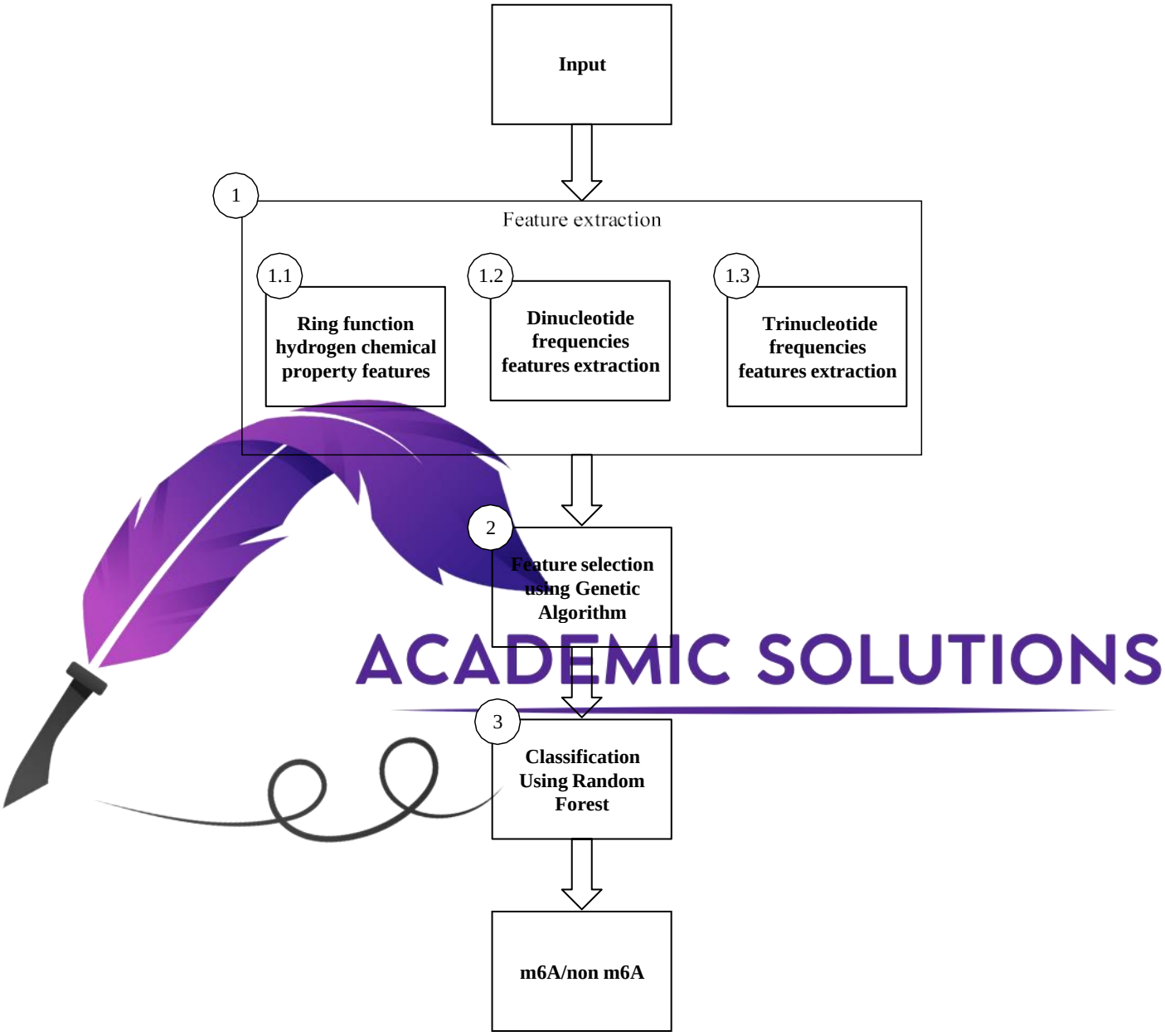


Figure 3.1: The flow diagram of the *m6A-pred* model for the detection of m6A sites in yeast RNA transcriptomes..

3.2.2.1 Feature extraction

Feature extraction is an important step in which meaningful patterns are extracted from Omics data [34]- [35]. In this step, the input Omics sequence are encoded and then converted into numerical values. These numerical values are called features and are represented in the form a vector. It is an important step, in which biological samples are transformed into numerical vectors, such that each vector captures different aspects of the the encoded information surrounding the m6A sites. In the current study, we construct features using fusion of three types of information based on the sequence characteristics, including, statistical, and chemical properties of the nucleotides, to precisely predict the presence of m6A sites in RNA sequences. The features are broadly classified into three types, namely, *Ring function hydrogen chemical property features*, *Dinucleotide composition features* and *Trinucleotide composition* features. A brief explanation of each type of feature are given bellow:

Ring function hydrogen chemical property features (RFHC): Inspired from its favorable results in capturing RNA's post-transcriptional modifications, we also utilized chemical property based nucleotide composition (also known as Ring function hydrogen chemical property) features to convert the input RNA sequence into a numerical feature vector [36, 37].

The input RNA sequence have four base pairs, namely, adenine, guanine, cytosine and uracil. These bases are represented by A, C, G and U, respectively. Each base have a different kind of chemical property associated with it. For example, if we consider the ring structure; the adenine and guanine belong to purines (having double ring structures), while cytosine and uracil are called pyrimidines (having a single ring structure). Likewise, for the second chemical property, if we consider the secondary structures of these nucleotides, there is strong hydrogen bond between guanine and cytosine, while weak hydrogen bond is present between adenine and uracil. Similarly, chemical grouping may be considered as the third property, in which, adenine and cytosine are members of amino group, whereas guanine and uracil are members of keto group.

To encode the input RNA sequence we use three coordinates (x, y and z) system. The

x, y and z represent the three chemical properties in the input RNA sequence. For each representation, we use the binary values, assigned to four nucleotides on the basis of these chemical properties [36, 38]. As an example, the ring structure is represented by x coordinates, hydrogen bonds based grouping is represented by y coordinate, and chemical grouping is represented by z coordinate. The three coordinates system (x_i, y_i, z_i) is used to represent each RNA nucleotide and it can be formulated using the following equation:

$$xi = \begin{cases} 1 & if\ si \in \{A, G\} \\ 0 & if\ si \in \{C, U\} \end{cases}$$

$$yi = \begin{cases} 1 & if\ si \in \{A, U\} \\ 0 & if\ si \in \{C, G\} \end{cases}$$

$$zi = \begin{cases} 1 & if\ si \in \{A, C\} \\ 0 & if\ si \in \{G, U\} \end{cases}$$

As an example, the nucleotide A in the sequence is represented by (1, 1, 1), C is represented by (0, 0, 1), G is represented by (1, 0, 0), and U is represented by (0, 1, 0).

In addition to previous three features, we also construct another feature called the nucleotide density composition, computed from the input RNA sequence [39]. The nucleotide density is represented by d_i of a nucleotide q_i at position i and is computed from the input RNA sequence using the following formula:

$$di = \frac{1}{|Ni|} \sum_{n=1}^{|Ni|} f_n(q_i);\ where\ f_n(q_i) = \begin{cases} 1; & if\ q_i\ is\ present\ at\ location\ n \\ 0; & otherwise. \end{cases} \quad (3.1)$$

Where, $|Ni|$ denotes sequence length of the *ith* prefix substring n_1, n_2,..., n_i in the input RNA sequence, and q_i is the nucleotide at location i and $q_i \in$(A, C, G, U), while $f_n(q_i)$ is the frequency of nucleotide q_i at location n.

For an input RNA sequence of length l, the RFHC feature results in a $(4 \times l)$-dimensional feature vector. In our case, each RNA transcript is 51 base pairs long, resulting in a

feature vector of size 4× 51, having a total of 204 values. For clarity, an example of RNA sequence encoding using the RFHC features is presented in Figure. 3.2.

Dinucleotide composition (DNC): Motivated by DNC success in RNA classification and DNA classification we also encode our input sequences using the Dinucleotide composition (DNC) technique [40, 41]. An RNA sequence R having L length long can be represented as:

$$R = [B_1B_2B_3B_4B_5B_6B_7B_8B_9B_{10}\ldots B_L] \tag{3.2}$$

where B_1 represents the first position residue of RNA sequence, B_2 represents the second position residue and B_L represents the L_{th} position residue in the RNA sequence. In dinucleotide composition (DNC), the occurrence frequency is computed on the base of pair of nucleotides [42]. The dinucleotide composition can be formulated as:

$$R = [f(AA), f(AC), f(AG), f(AU), \ldots, f(UU)]^T \tag{3.3}$$

Where, f(AA) denotes the number of occurrences of AA pair, f(AC) denotes the number of occurrences of AC pair, f(AU) is the number of occurrence of the pair AU and so on up to f(UU) which represents the number of occurrences of the pair UU. This step gives a total of 4^2 (i.e., 16) dimensional feature vector.

Trinucleotide composition (TNC): Motivated by Trinucleotide Composition (TNC) success in RNA classification and DNA classification we also encode our input sequences using this technique [40, 41]. In trinculeotide compostion three bases are combined as single unit and the frequency of each unit is counted. The trinucleotide composition is formulated as:

$$R = [f(AAA), f(AAC), f(AAG),\ldots, f(UUU)]^T \tag{3.4}$$

$$R = [f(1), f(2), f(3), f(4), f(5), f(6)\ldots, f(64)]^T \tag{3.5}$$

Where the term f(AAA) is number of occurrences of AAA bases, f(AAC) is total number the occurrences of AAC bases, f(AAG) is the total number of occurrences of AAG and

RNA sequence	A	C	G	U	G	A	C	G
Ring structure	1	0	1	0	1	1	0	1
Hydrogen bond	1	0	0	1	0	1	0	0
Functionality	1	1	0	0	0	1	1	0
Nucleotide density composition (di)	1	0.5	0.33	0.25	0.4	0.33	0.29	0.38
Feature vector-{1,1,1,1,0,0,1,0.5,1,0,0,0.33,0,1,0,0.25,1,0,0,0.4,1,1,1,0.33,0,0,1,0.29,1,0,0,0.38}								

Figure 3.2: The encoding scheme for m6A sites using nucleotide chemical properties and nucleotide composition

f(UUU) is total number of occurrences of UUU in the input RNA sequence. The corresponding feature vector will be 64 dimensional feature vector. At the end of this step, three different type of vectors, i.e., *Ring function hydrogen chemical property features*, *Dinucleotide composition features* and *Trinucleotide composition* features are combined to make a single high dimensional feature vector. The feature vector contains a total of 284 features for each input RNA sequence.

3.2.2.2 Feature selection

Feature selection is an important preprocessing step in which subset of features are selected from the whole set of features to construct a model. There are several reasons for feature selection, including simplifying model for researcher's interpretation, decreasing model training time, avoiding curse of dimensionality, or enhancing generalization to reduce overfitting [43]. Different types of feature selection methods have been proposed to tackle this problem [44]-[45]. Inspired from the works of [46, 47], we chose Genetic Algorithm (GA) for feature selection.

The genetic algorithm is a type of evolutionary algorithms. It is a metaheuristic algorithm that works on the principle of natural selection. It produces efficient results for optimization and searching problems and is also favored due to its parallel nature. In classification

problem, right number of variable are picked up by genetic to make a predictive model. The main advantage of GA on other feature selection techniques is, it produce best solutions from earlier best solutions. Therefore, selection is improved over time. The working principle of GA is to extract best genes (variables) by combining different solutions from generation after generation. Hence, it produces fitted individuals. Variables inside every possible solution are considered as whole not a single entity and no single variable is ranked against the target. Therefore, we already familiar with the fact, that variables work well in group [48]. Figure. 3.3 shows flowchart of the GA algorithm.

For the high dimensional feature vector from the previous step, we construct the population of our algorithm. The individuals (also called chromosomes) in our populations are binary vectors of length 284. We start with a population of 50 individuals and iterate using generational genetic algorithm. Each bit (or gene) in the chromosome represents the presence or absence of a particular feature. Thus, the gene value 1 means that the feature is present, while gene value zero represent absence of a particular feature.

In the second step we define the fitness function of our individuals in the population. For fitness function we use the accuracy metric. We select the individuals on the basis of their fitness. Higher the fitness, higher is the selection probability and vice versa. Once the pairs are selected, we then apply the single point crossover operator. This operator allows convergence of the solution towards the maxima (which can be either local maxima or global maxima). To avoid local maxima, the algorithm should do exploration of the search space. For exploration, we utilized the mutation operator, which randomly flips 0s to 1s and 1s to 0s with a probability of 0.01.

The GA algorithm works iteratively until fitness value improves the objective function criteria. For our problem we executed the GA for 1000 iterations for which the fitness of the population started to converge towards the best solution. At the end, the genetic algorithm resulted the best individual with 39 optimal features. This optimized feature vector is given as input to a classifier in the next step.

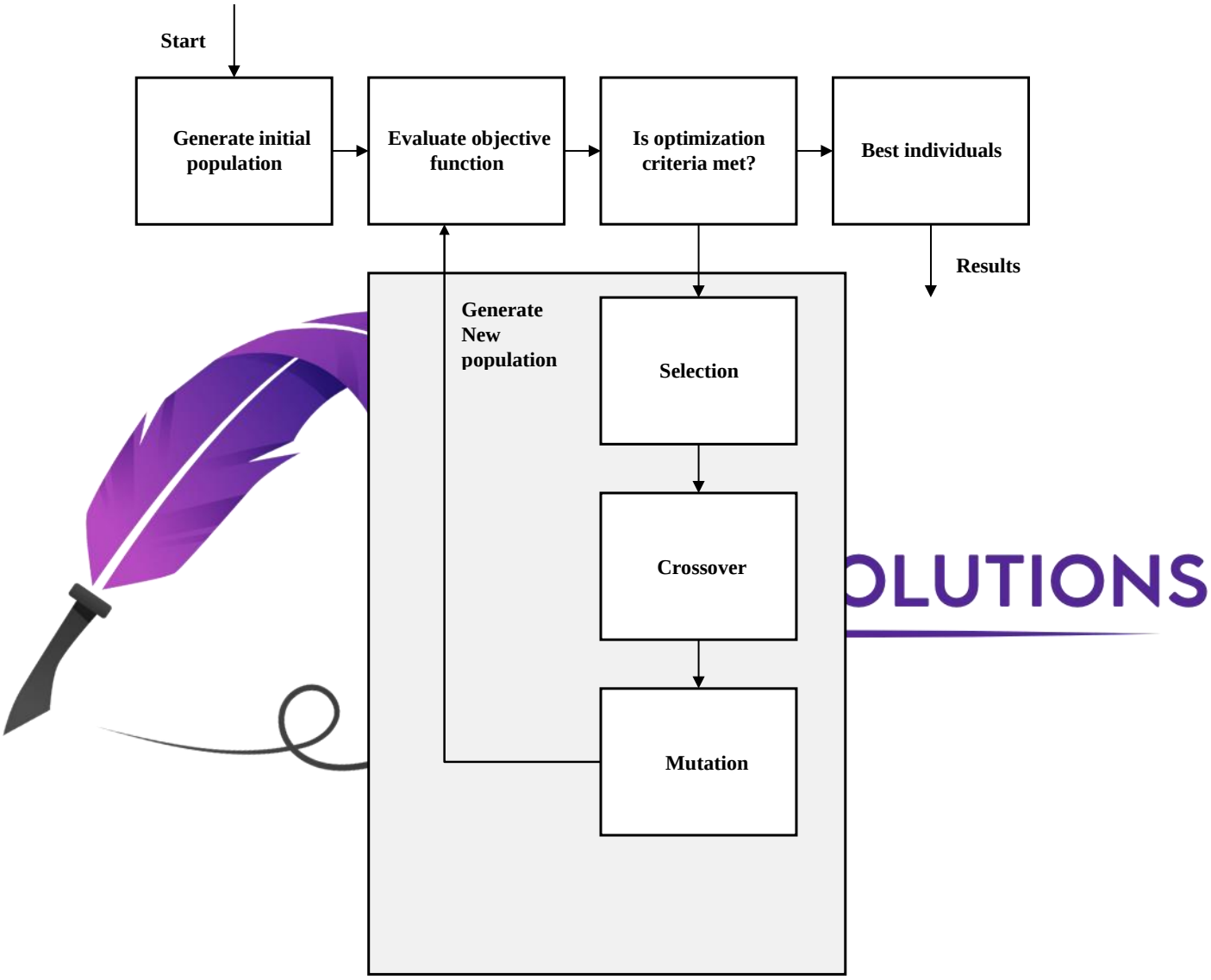


Figure 3.3: Framework of Genetic Algorithm

3.2.2.3 Classification

We employed random forest algorithm (a type of ensemble classifier) to separate m6A sites from non-m6A sites. The ensemble classifiers utilize training set efficiently that leads to improve model accuracy and generalization [49, 50, 51, 52]. Random forest work well on high dimension data, because it make forest of trees which produces lower bias and final prediction is based on majority of voting therefore, it reduces model variance. Another beautiful property of Random forest is,it improves its model by selecting and ranking variables according to their importance. Hence, these properties of random forest make it best fitted classifier for genomic data [53]. We applied bagging and random feature selection techniques for classification. In the bagging step, bootstrap training sample is used to build up the trees and final prediction is made using majority counting. We constructed 500 trees from the bootstrap samples and further use two variables for tree splitting. The final decision is made based on the 500 trees model along with majority voting.

3.3 Experimental Setup and Results

3.3.1 Dataset

The proposed *m6A-pred* predictor is evaluated on benchmark dataset taken from [13]. The dataset contains 1307 sequences containing m6A sites and 1307 sequences without m6A sites. The proposed predictor takes a sample as input, extracts three type of features from that sample, then reduces feature using GA algorithm and final 39 feature vector is given input to random forest to predictor whether this is an m6A site or non m6A site.

3.3.2 Jackknife validation

The predictor's performance is measured on the basis of its prediction quality. Thus, it is crucial to evaluate its prediction quality in accurate way. In this paper, we test our predictor quality on the basis of Jackknife testing technique. In this technique, one sample

is selected for testing and the remaining samples are used for training. The process is repeated *n* times, where *n* is the number of instances in the dataset. Thus, the total number of iterations are equal to the number of samples in dataset. The results are averaged over all iterations.

3.3.3 Evaluation metrics

The problem tackled in this article have two classes i.e. binary classification task. In present study our classification algorithm will predict a sequence as a m6A site or non-m6A site. There are different metrics which can be used for comparison but for binary classification four metrics are commonly used for measuring the prediction quality of classifier. These metrics are sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC) and accuracy (Acc). The mathematical formulations of the above metrics are given bellow:

$$Sn = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{3.6}$$

Where TP, TN, FP and FN stand for true positives, true negatives, false positives and false negatives, respectively. Sensitivity is formulated in equation which avoids false negative. It ranges from zero to one. If Sn=1 then none of the m6A sites in positive subset is predicted false negative. On other hand if Sn=0 that mean all of m6A sites are predicted as false negative i.e. members of non-m6A sites class.

$$Sp = \frac{TN}{\mathrm{TN} + \mathrm{FP}} \tag{3.7}$$

Similarly specificity avoids false positive rate and also ranges from zero to one. The Sp=1 means all non-m6A sequences are predicted as non-m6A or in other word none of them are classified as m6A sites. On other side Sp=0 means all of non-m6A sites are classified incorrectly.

$$Acc = \frac{TP + TN}{\mathrm{TP+TN+FP+FN}} \tag{3.8}$$

Accuracy avoids both false negatives and false positive. Accuracy measures the total number of samples predicted accurately divide by the total number of samples in the data

set. Accuracy ranges from zero to 1. The zero value of accuracy means that all samples are predicted incorrectly, while Acc=1 means none of the sample in both classes are predicted incorrectly. Accuracy is not considered a good metric for imbalance dataset.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.9}$$

Another important metric called Matthews correlation coefficient is used to check the quality of predictor. It will be a best choice if the data is imbalance. MCC has a lower limit of -1 and an upper limit of +1. When MCC=-1 that means all of samples in both classes are predicted incorrectly. When MCC=0 then the predictions are like a random guess. When MCC=1 that means all samples in both classes are predicted correctly.

3.3.4 Performance of *m6A-pred* on combined features

The result of *m6A-pred* predictor using all features combined (i.e., 284 features) are reported in Table. 3.1. It can clearly be seen that the predictor has high sensitivity value, which is due to that the fact that the the predictions have high number of true positives while comparatively very less number of false negatives. Likewise, the specificity is also high which means that the predictions have high number of true negatives while less number of false positives. The predictor also has acceptable value for accuracy and Matthews correlation coefficient. However, the accuracy value is less than sensitivity and specificity values. The reason for this slight reduction in accuracy are the number of false positives, which are slightly higher than the number of fasle negatives.

It is pertinent to mention here that for the m6A site detection problem, the Receiver Operating Characteristics (ROC) curves for existing predictors are not reported in literature. However, for ease in future comparisons, we also show the ROC curve for the *m6A-pred* predictor in Figure. 3.4. The *m6A-pred* achieved an AUC score of 0.82.

3.3.5 Genetic algorithm for feature selection

In a classification problem, selecting the right kind and right number of features is of pivotal importance. For this purpose, we employed the state of the art genetic algorithm (GA) technique in our problem. The main advantage of GAs on other feature selection

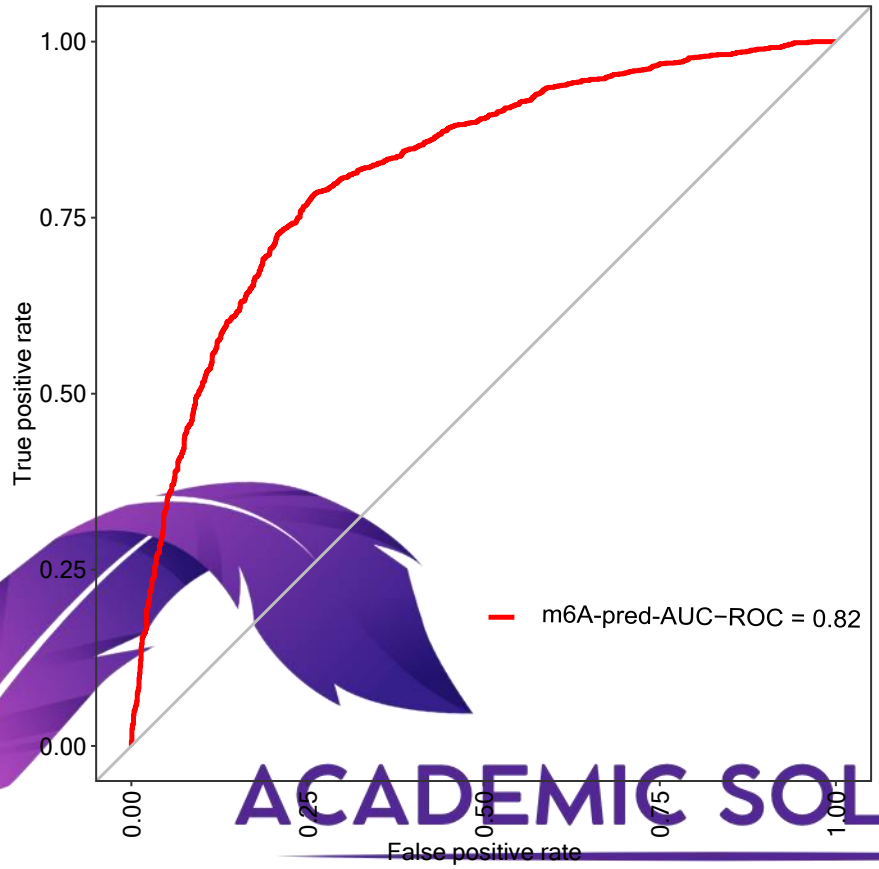


Figure 3.4: The ROC (Receiver Operating Characteristics) for *m6A-pred* predictor.

Table 3.1: Performance of the m6A-Pred on all feature types combined

Feature	Classifier	sn(%)	sp(%)	Acc(%)	MCC(%)
All 284 features (without feature selection)	KNN	75.74	51.26	63.50	27.86
All 284 features (without feature selection)	SVM	75.74	76.00	75.78	51.56
All 284 features (without feature selection)	Random forest	**77.89**	75.59	**76.74**	**53.84**

Table 3.2: Comparison of different feature selection techniques using random forest

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC(%)	# of Features
Random forest-MRMR	76.43	76.59	76.51	53.02	50
Random forest-SA	72.38	74.45	73.41	46.83	126
Random forest-SBF	75.21	76.43	75.82	51.64	126
Random forest-ACO	74.98	70.39	72.69	45.41	226
Random forest-EFS	65.18	63.31	64.50	29.00	50
Random forest-WOA	75.21	74.90	75.06	50.11	145
Random forest-RFE	78.34	77.35	77.85	55.70	39
Random forest-GA(m6A-Pred)	**79.65**	**77.51**	**78.58**	**57.17**	**39**

techniques is through its genetic operators (specifically, crossover and mutation operators) that always try to optimize the solution state i.e., it always tries to produce a better solution from previous solutions. Therefore, fitness of solutions is improved over time. We chose GAs for feature selection because of this property as well as its speed to find the optimal solution due to its parallel nature.

To justify our choice of GAs, we in table 2, present results of different state of the art feature selection algorithms when applied to the same problem using Random forest classifier. It can clearly be seen that the GAs stand out among all the feature selection algorithms with optimum number of features, while producing the best values for all the statistical measures including, sensitivity, specificity, accuracy and Matthew correlation coefficient. The GA also outperforms famous feature selection algorithms namely, recursive feature elimination and Minimum Redundancy Maximum Relevance (mRMR). The main reason behind this gain is the selection of the most discriminative set of features by the GA. In addition, it also selects the optimum number of features which results in reduced time complexity of the algorithm.

3.3.6 Performance of our predictor on reduced features

As addressed in section 2.4, we have also reduced our feature space by using the genetic algorithm feature selection technique. We used the random forest and some of other classifiers to judge the performance of selected features. The results in table 2 show that random forest perform well on selected features compared to other techniques. If we compare the results of Table 3.1 with last row of Table 3.2, it is clear that performance of

Table 3.3: Performance of the *m6A-Pred* on reduced feature types

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC
SVM-MRMR	76.81	75.98	76.40	0.5279
Random forest-MRMR	76.43	76.59	76.51	0.5302
SVM-RFE	77.96	77.27	77.62	0.5524
Random forest-RFE	78.34	77.35	77.85	0.5570
SVM-GA	78.11	77.12	77.62	0.5524
m6A-pred	**79.65**	**77.51**	**78.58**	**0.5717**

selected features is better than combined features. These results also show that the feature selection play a vital role to improve the accuracy as well as the time complexity.

3.3.7 Statistical Analysis

The 5x2cv paired *t-test* is a method used to compare the performance of two prediction models. It was proposed by [54] to tackle limitations of other techniques such as the resampled paired *t-test* and the k-fold cross-validated paired *t-test*.

To explain this technique, we take two classifiers named A and B. Furthermore we have a dataset, in which all instances have labels. In common holdout validation technique, we split the dataset into two parts: a test set and a training set. In 5x2cv paired *t-test*, we split the dataset in 50% test and 50% training and iteratively repeat it five times.

In the next phase, classifier A and B are trained on the training set. The performances of classifiers A ($\mathbf{p_A}$) and B ($\mathbf{p_B}$) are evaluated on the test set. Then we rotate the training set and test set i.e. training set becomes test and vice versa and classifiers are evaluated again. The two performance differences of these classifiers is mathematically described below.

$$p^{(1)} = p_A{}^{(1)} - p_B{}^{(1)} \tag{3.10}$$

&

$$p^{(2)} = p_A{}^{(2)} - p_B{}^{(2)} \tag{3.11}$$

Next, we compute the mean and variance based on these differences:

$$\bar{p} = \frac{p^{(1)} + p^{(2)}}{2} \tag{3.12}$$

and

$$s^2 = (p^{(1)} - \bar{p})^2 + (p^{(2)} - \bar{p})^2 \tag{3.13}$$

Next on the bases of five iterations variance-of-the-differences, we finally calculate *t* statistic as follow.

$$t = \frac{p_1^{(1)}}{\sqrt{(1/5)\sum_{i=1}^{5} s_i^2}} \tag{3.14}$$

Where $p_1^{(1)}$ is p_1 calculated in the first iteration. While in *t* statistic, we assume that, it follows the *t*-distribution with 5 degrees of freedom and under null hypothesis both of the models have equal performances. On the bases of *t* statistic we calculate p and compare it with significance level i.e., α =0.05. If the calculated **p** value is less than 0.05, we reject null hypothesis and accept that there is significant difference between two models.

In our example, we applied 5x2cv paired *t-test* method to check weather there is significant differece between model constructed based RFE and GA selected features. *t* statistic: -5.385 and **p**-value: 0.003 shows that, our predictor m6A-Pred outperform Random forest-RFE.

3.3.8 Why GA performs better?

In a classification problem, the right number of variables are picked up by the genetic algorithm to make a predictive model. The main advantage of GA on other feature selection techniques is, it produces the best solutions from earlier best solutions. Therefore, selection is improved over time. The working principle of GA is to extract the best genes (variables) by combining different solutions from generation after generation. Hence it produces fitted individuals. Variables inside every possible solution are considered as a whole, not a single entity and no single variable is ranked against the target. Therefore, we already familiar with the fact that variables work well in a group.

If we look at a table, the genetic algorithm outperform famous feature selection algorithms: recursive feature elimination and MRMR. The main reason behinds that it selects features that have more discriminative power than features selected by other algorithms. Besides that, Genetic algorithms reduced time complexity due to less number of features.

Table 3.4: Performance of the m6A-Pred on all feature types combined

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC
pRNAm_PC[19]	69.72	69.75	69.74	0.40
M6A_HPCS[23]	77.35	67.41	72.38	0.45
RAM-ESVM[24]	78.93	**77.78**	78.35	0.57
imethyl- STTNC[25]	70.32	68.17	69.84	0.38
M6AMRFS[14]	75.21	73.39	74.25	0.4852
m6A-pred	**79.65**	77.51	**78.58**	**0.5717**

3.3.9 Comparison of m6A-pred with existing approaches

In this section, we compare our method with state of the art methods that are reported to produce good results for the same type of problems. As can be seen in Table 3.3, four commonly used validation metrics are picked for comparison. We compare the results of our predictor to currently developed predictors i.e. *pRNAm_PC*, *M6A_HPCS*, *RAM-ESVM*, *imethyl-STTNC*, *and M6AMRFS*.

Same benchmark dataset and same validation (i.e., jackknife validation) are used to keep fairness in comparison. As shown in Table 3.3, the *m6A-pred* predictor is superior in performance than all of the recently proposed predictors. Except specificity values, the *m6A-pred* produced higher values for sensitivity, accuracy and Matthews correlation coefficient. The *m6A-pred* outperforms pRNAm_PC, M6A_HPCS , imethyl-STTNC, M6AMRFS in all four metrics. The *m6A-pred* outperforms RAM-ESVM in all metrics except for specificity in which RAM-ESVM better performs. The reason for this higher value lies in slightly higher number of false positives for *m6A-pred*, while *m6A-pred* is much better in terms of the number of false negatives.

3.4 Discussion

N6-methyladenosine (m6A) is frequently occurring post-transcriptional modification on both eukaryote and prokaryote messenger RNA transcripts. The regulation of messenger RNA is facilitated by RNA binding proteins which recognize these m6A modification sites. The occurrence of m6A on small portion of transcript performs vital roles in cellular function like cellular-stability, pre-mRNA splicing, microRNA regulation and some

other important biological functions, thus indicating a link with various biological processes and diseases. In this work, we propose a novel method (called *m6A-pred* predictor) that utilizes a fusion of characteristics including, statistical, and chemical properties of the nucleotides, to precisely predict the presence of m6A sites in RNA sequences. The *m6A-pred* predictor overcomes the problem of inefficient detection of m6A sites in yeast transcriptomes (due to varied structure) and inability of the computational techniques to capture the encoded information surrounding the m6A sites.

The *m6A-pred* utilizes a variety of features and fuses them to capture the encoded information surrounding the m6A sites in ordrer to efficiently detect the m6A sites in yeast transcriptomes. The extracted hybrid features are high dimensional and capture different aspects of the m6A sites. In the next subsection, we present a brief analysis on how feature fusion improves the m6A site prediction.

3.4.1 Analysis of *m6A-pred* performance using feature fusion

For the analysis, we examine the performance of our predictor on both individual feature representations as well as their fusions. Three different types of feature encoding methods are used for numerical representation of the input sequences. The three feature types (also called feature sets) make a total of seven unique combinations. In order to perform analysis using these unique combinations, we pick four widely used classification metrics, i.e., sensitivity, specificity, accuracy and Matthews correlation coefficient, as reported in Table 3.5.

By comparing feature FS_1 and FS_2 in Table 3.5, it can be seen that trinucleotide features have more discrimination power than dinucleotide features. Likewise, if we look at FS_3 in Table 3.5, the position specific chemical property based features are more powerful than the previous two features. An interesting observation can be seen in 3.5, which shows that by integrating FS_1 with FS_2 and FS_3 degrades the performance of trinucleotide based features and position specific chemical property based features because dinucleotide may have some features which have low discrimination power.

In Table 3.5 it can be seen that adding FS_2 with FS_2 significantly increases the perfor-

Table 3.5: Performance of *m6A-Pred* on individual feature types as well as their fusion

S#	Feature Sets	Predictor	Sn(%)	Sp(%)	Acc(%)	MCC(%)
1	FS_1	Random forest	66.60	63.80	65.20	0.3046
2	FS_2	Random forest	67.33	68.85	68.10	0.3619
3	FS_3	Random forest	75.90	74.75	75.32	0.5065
4	FS_1+FS_2	Random forest	66.80	63.40	65.10	0.3023
5	FS_1+FS_3	Random forest	75.51	74.21	74.86	0.4701
6	FS_2+FS_3	Random forest	76.97	74.60	75.79	0.5158
7	FS_1+FS_2+FS_3	Random forest	**77.89**	**75.59**	**76.74**	**0.5384**

FS_1: Dinucleotide composition
FS_2: Trinucleotide composition
FS_3: Ring function hydrogen chemical property

mance of classifier for all four metrics. Lastly, FS_1+FS_2+FS_3 indicates that adding all three types of features further enhances the performance of the classifier. This shows that the fusion of features produces most discriminating features for separating m6A from non-m6A sites. At the end, it is concluded that feature fusion plays a vital role to enhance the performance of predictor.

Chapter 4

m6A-Finder: Detecting m6A Methylation Sites from RNA Transcripts using Feature Fusion

4.1 Introduction

Recently various types of post-transcription modifications have been discovered in messenger RNA (mRNA) [55]. The first discovered amongst them was m6A, which plays a vital role in several biological processes. It performs an essential role in cellular processes like mRNA splicing and stability [32], localization and degradation of RNA [3], and microRNA (miRNA) regulation [56]. Researchers reported that any abnormality in m6A could cause life-threatening diseases like breast cancer [4], leukemia [5], prostate cancer [6], thyroid tumor [7], etc. Therefore, correct identification of m6A sites would be beneficial for both basic research and personalized-medicine.

Researchers tried high-throughput wet lab techniques like "high-performance liquid chromatography" [15], "thin layer two-dimensional chromatography" [16] and next-generation sequencing technique [12] for identification of m6A sites in the genome. However, these techniques are suffering from some problems such as their results are not so good, have high time complexity and are costly. The mentioned restrictions of these techniques compelled the researchers to develop machine learning-based computational tools to identify

m6A sites in the genome. Recently, some computational predictors have been developed to identify m6A sites in different species like "Saccharomyces cerevisiae" [13], "Homo sapiens" [33] and "Arabidopsis thaliana" [57]. Among all these species, the Saccharomyces cerevisiae is the most explored one for research purposes.

Recently researchers proposed two predictors named m6Apred [18] and pRNAm-PC [19] to detect m6A site in Saccharomyces cerevisiae transcriptome. The first one uses chemical properties based feature extraction scheme while second uses chemical properties based auto-covariance for feature extraction. The first one has low sensitivity while the second one has low accuracy. The main reason behind their low performance is that the feature extraction schemes they used are not sufficient to discriminate m6A sites from non-m6A sites.

To identify m6A sites in the cross-species, SRAMP [20] was proposed. The authors of the predictor used binary encoding scheme for feature representation. The said predictor achieved reasonable specificity, but the sensitivity was very low. Similarly to SRAMP, Chen et al. proposed methyRNA [21] to identify m6A sites in cross-species. The authors used nucleotide chemical properties and density information for feature extraction. The predictor showed promising results on mammalian data, but for Saccharomyces cerevisiae transcriptome, its accuracy was low.

Likewise, the above-stated predictors, researchers proposed methyl-STTNC [25] via using split tri-nucleotide and tetra-nucleotide frequency composition to find m6A sites in multiple species data. The predictor performed well on all datasets except Saccharomyces cerevisiae. Inspired by [25], some researchers developed a predictor named BERMP [26] to find m6A sites for the same problem using Nucleotide composition. They used XGboost for classification. The BERMP faced the same problem as methyl-STTNC.

Similar to methyl-STTNC, another predictor named m6A-MRFS [14] was introduced to detect m6A sites in multiple species data. It extracts features based on position-specific dinucleotide computation and binary encoding. The mentioned predictor performed well on all datasets except yeast. Meanwhile, another predictor "iN6-Methyl" [58] was proposed to detect m6A sites in different species. Researchers of this predictor used word2vec as a feature extraction scheme and CNN for classification. The proposed

predictors has two downsides. Its complexity is high due to deep learning classification and its accuracy is low compared to other predictors.

In recent times, a predictor known iRNA-Freq [29] was introduced to detect m6A site in yeast. Authors of the mentioned paper used frequent gap-Kmer approach for feature extraction. The linear regression based classifier is utilized for prediction. The specificity of the predictor was acceptable but its sensitivity was low as compared to other predictors. Most recently, a predictor called iMRM [30] was proposed to detect m6A sites in cross-species data. The iMRM uses hybrid feature extraction technique based on chemical and statistical facets of RNA. The iMRM achieved encouraging results on all datasets except Saccharomyces cerevisiae.

From the intensive review, it is concluded that feature extraction has an essential role in the performance of a predictor. The mentioned feature representation techniques in literature are not enough to discriminate m6A sites sequences from non-m6A site sequences. Besides chemical and statistical features, another facet of RNA which is the physical property has not been explored yet. In this research, we propose a novel predictor named *m6A-Finder* which uses Hexa-nucleotide composition with auto-correlation features, based on physical properties to capture hidden details surrounding m6A sites. The proposed predictor capture both local and global sequence order information therefore, it will resolve the problem faced by existing predictors in the identification m6A sites in Saccharomyces cerevisiae due to non-discriminative features. In this research, local Hexa-nucleotide composition increases the vector dimension which causes overfitting. To solve this problem, we use a filter-based feature selection techniques to select optimal features and reduce time complexity.

The key contributions of this study are as follow:

1. Novel hexa-nucleotide composition features with autocorrelation function (ACF) features based on physical properties, are extracted from input sequences.
2. Minimum Redundancy Maximum Relevance, i.e, mRMR algorithm is used to select optimal features and reduce time complexity.

The rest of the paper is structured as follows: we explain the proposed methodology of

the *m6A-finder* in Section 2. In Section 3, we present our results on RNA transcript of Saccharomyces cerevisiae species. The section 4 presents the detailed discussion of how the proposed method overcomes the highlighted problems. Finally, we conclude our study in Section 5.

4.2 Materials and Methods

4.2.1 Benchmark Dataset

The Saccharomyces cerevisiae genome share a consensus motif GAC in which the center motif has the potential to be methylated. This is very well represented by the dataset prepared by Chen [13], and we use it as a benchmark to evaluate the performance of our predictor. The dataset is composed of a total of 2,614 sequences, in which half samples (i.e., 1,307 transcripts) contain experimentally confirmed m6A sites, while the other half samples doesn't contain the m6A sites. All of the instances have identical length i.e., 51bp, and none of them have a similarity higher than 85%.

4.2.2 The *m6A-Finder* Model

The state-of-art predictors in the existing literature showed low performance on yeast transcriptomes due to intricate arrangement of residues around the m6A methylation sites. The fundamental problem lies with the feature representation schemes utilized to detect the m6A sites, that are unable to efficiently distinguish the m6A sites from non-m6A sites. To handle this problem, we developed a novel model called the m6A-Finder based on statistical features (i.e., Hexa-nucletotide composition of residues) and physical features (i.e., autocorrelation features based on eight physical properties of residues) to predict the presence or absence of m6A sites in the input RNA transcript. The overall scheme is presented in Figure 4.1. The m6A-Finder consists of four steps, namely, 1) Statistical Feature Extraction 2) Physical Feature Extraction 3) Feature Selection 4) Classification. The detailed explanation of each of the preceding step is given below:

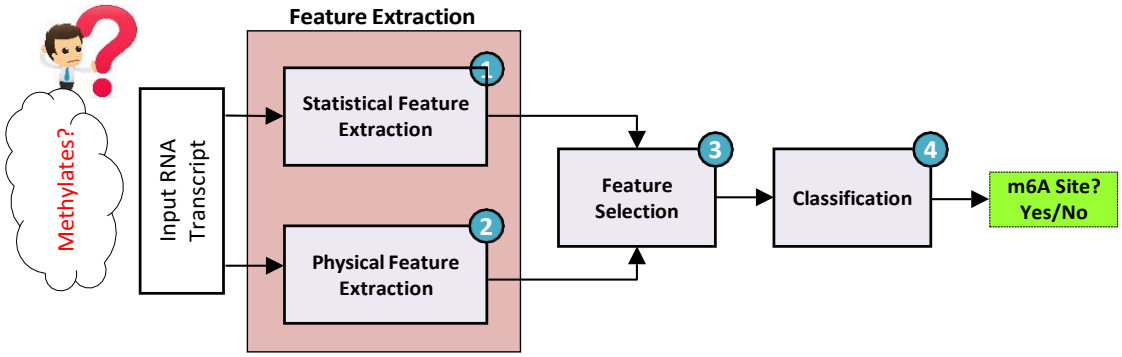


Figure 4.1: The diagrammatic representation of the m6A-Finder for identification of m6A site in Saccharomyces cerevisiae transcriptome.

4.2.2.1 Statistical Feature Extraction

Feature extraction is one of the key step that affects the classification accuracy. In this step, we derive useful patterns from the input data which are sometimes not so trivial. For our problem, the first type of features that we extract are based on the statistical properties of the nucleotides in the input RNA sequence. Statistical properties are of paramount importance because they help capture the local sequence order present in the input RNA sequence. For this purpose, we took inspiration from the seminal work of *Balachandran Manavalan* [59], which gave encouraging performance of nucleotide composition for DNA sequences. Inspired from this work, we also incorporate hexa-nuceleotide composition for the extraction of statistical features. In hexa-nucleotide composition technique, the frequencies of six adjacent nucleotides are calculated. For an RNA sequence, there will be a total of 4^6 (i.e., 4096) unique hexa-nuceleotide features, where 4 represents the number of unique bases and 6 represents length of a unique feature. These features are mathematically represented by the following equation:

$$R = [f(AAAAAA), f(AAAAAC), f(AAAAAG), f(AAAAAU), \ldots, f(UUUUUU)]^T \quad (4.1)$$

where, R represents a feature vector, *f(AAAAAA)* is the frequency of six adjacent nucleotides i.e., AAAAAA, while *f(AAAAAC)* is the frequency of another six adjacent nucleotides i.e., AAAAAC, and so on. For the m6A methylation problem the hexa-nucleotide composition outperforms all other N-gram based statistical features.

Table 4.40: List of physical properties associated with each

Physical property	A	C	G	U
Density	2.1	1.6	2.2	1.5
Boiling point	676.3	445.8	561.5	440.5
Melting point	234	325	360	330
Flash point	362.8	223.4	293.4	220.02
Mass	267	111.04	151.04	112.02
PSA	139.54	71.77	100.45	65.72
logP	-1.02	-2.29	-2.03	-2.55
Index of Refraction	1.907	1.689	2.047	1.64

4.2.2.2 Physical Feature Extraction

The hexa-nucleotide composition features only capture the local sequence order information, while ignoring the global order present in the input sequence. In order to capture the global sequence order information, we calculate the physical property based features. For this purpose, we chose eight novel physical properties of individual nucleotides i.e., density, boiling point, melting point, flash point, mass, Polar surface Area (PSA), Partition coefficient (logP), and index of refraction. The details of physical property based values for each nucleotide can be found in Table 5.1.

4.2.2.3 Feature extraction

Feature extraction is one of the essential step for accurate classification of genomic data. In the feature extraction step, we derive hidden useful patterns from data. First of all, we convert the genomic sequence into numerical values. Then, we extract features from the genomic sample on the base of these numerical values. These features make the feature vector. All attributes inside this vector make vector dimension. For the current problem, we used Hexa-nucleotide composition with autocorrelation features based on physical aspects of nucleotides. These diverse features have both local order and global order information. The detail of these features is given bellow: The feature vector is constructed by modeling similarity between physical properties in the input sequence over lagged version by itself. For this purpose, we encode the physical property values

Table 4.2: List of physical properties associated with each nucleotide.

Physical property	A	C	G	U
Density	2.1	1.6	2.2	1.5
Boiling point	676.3	445.8	561.5	440.5
Melting point	234	325	360	330
Flash point	362.8	223.4	293.4	220.02
Mass	267	111.04	151.04	112.02
PSA	139.54	71.77	100.45	65.72
logP	-1.02	-2.29	-2.03	-2.55
Index of Refraction	1.907	1.689	2.047	1.64

into a feature vector using the autocorrelation function (ACF) as shown in Equation 4.2.

$$ACF(P_j, k) = \frac{1}{(n-k)\sigma^2} \sum_{i=1}^{n-k} (R_i^{P_j} - \mu)(R_{i+1}^{P_j} - \mu) \tag{4.2}$$

where, P_j is jth physical property value with $j \in [1, 2, 3, \ldots, 8]$, k is the lag value for auto-correlation, with $k \in [1, 2, 3, \ldots, 15]$, n is the length of the input sequence, μ is the mean value of the input sequence, σ^2 is the variance of the input sequence, and $R_i^{P_j}$ is the value of residue R_i with property P_j.

In order to keep values separated for each lag, the values obtained for each property from Equation 4.2 are averaged for each lag using Equation 5.4.

$$F_{physical}(k) = \frac{1}{T} \sum_{j=1}^{T} ACF(P_j, k) \tag{4.3}$$

where, T represents the total number of physical properties (eight in our case), and $F_{physical}(k)$ represents a feature value for each lag k. This step results in a total of 15 feature values for fifteen lags. We chose 15 after exhaustively experimenting with different lag values. Overall, both statistical and physical property based features result in a total of 4,111 features values. The combined feature vector resulted in an overfitting, therefore, to choose the optimum features set in the next step we apply the feature selection operation.

4.2.2.4 Feature selection

We have extracted features from RNA sequences which result in overfitting. To solve this problem, we use feature selection technique to include relevant features and exclude unnecessary ones. The key benefits of feature selection is that it reduces time complexity of the predictor and improves its prediction. There are two types of features selection techniques: wrapper based and filter-based. For our problem, the wrapper based methods suffer from overfitting and have high time complexity. Therefore, we use a filter-based approach i.e., Minimum Redundancy Maximum Relevance (mRMR). The mRMR method is proposed by Peng et al [60]. The mRMR uses a set of statistical functions to calculate relevance and redundancy among features. The core mechanism of mRMR is to select a feature at each iteration that has high relevance to the target class and minimum redundancy to previously selected features.

For example at iteration i, the score-$FCQ_i(f)$ for each feature is calculated by following formula:

$$score\_FCQ_i(f) = \frac{F(f|tc)}{\sum_{s \in f\,features selected\,until\,i-1} |cor(f,s)|/(i-1)} \quad (4.4)$$

The best feature at iteration i is the one having highest score, while score-$FCQ_i(f)$ is F-test with a correlation quotient. In numerator the function F is actually F-statistic, which calculates the relevance between feature f and target class (tc). Similarly, in denominator the function *cor* calculates redundancy between feature f and already selected features (s) using Pearson correlation.

The mRMR feature selection method selected 300 features from 4,111 features. These 300 optimal features have maximum relevance to target class and minimum redundancy with each other. After this step, these 300 optimal features go to support vector machine.

4.2.2.5 Classification using SVM

We utilized Support vector machine (SVM) to classify m6A sites from non-m6A sites [61]. It works on the basis of statistical theory to draw hyperplane between classes. To

improve the prediction, it maximizes the marginal distance from hyperplane [62]. It has multiple kernels to work with the different distribution of data [63]. In this work, we have used three kernels i.e., linear, RBF and polynomial. The Radial Basis Function (RBF) kernel produced good results that show our reduced features have non-linear distribution. The reduced vector of 300 features is given as input to SVM, it draws hyperplane according to RBF kernel function. Finally, it predicts the presence or absence of m6A site in a given RNA sequence.

4.3 Experimental Setup and Results

In this section, we discuss experimental setup and evaluation metrics in detail. Furthermore, we assess m6A-Finder performance on all feature and reduced feature sets. Finally, we compare our proposed predictor with existing state of the art models.

4.3.1 Jacknife validation

The prediction quality of a predictor depends on how correct it predicts the unseen instances. Therefore, it is crucial to validate predictor performance in the correct way. For the current study, we selected the Jackknife validation technique to check the performance of our predictor because it has low bias compared to hold-out and k-fold cross-validation. In Jackknife validation, a single sample goes as a test sample to predictor and the remaining are kept for training. The same process iterates for N number of times, where N is the total RNA samples in the dataset.

4.3.2 Performance metrics

Our problem belongs to the supervised binary classification task, in which predictor will identify whether the input RNA sequence has an m6A site or not. For performance assessment, we picked four commonly used metrics i.e., sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew correlation coefficient (MCC) [64]. The mathematical description and formulation of these metrics are given below:

The sensitivity measures how many samples in the m6A class are predicted correctly. The sensitivity ranges between zero and one. The zero sensitivity means that the predictor mispredicted all m6A sites in dataset. Similarly, a sensitivity value of one shows that the predictor accurately classified all of them. The mathematical formulation of sensitivity can be found in Equation 4.5.

$$Sn = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{4.5}$$

Likewise, specificity measures how many samples in non-m6A site class are predicted accurately. Like sensitivity, specificity also ranges between zero and one. The zero specificity shows that the predictor misclassified all sequences of the non-m6A class. The specificity value of one means the predictor classified all of them correctly. The mathematical description of sensitivity is given in Equation 4.6.

$$Sp = \frac{TN}{\mathrm{TN} + \mathrm{FP}} \tag{4.6}$$

The third statistical measure that we utilized is accuracy measure. The accuracy is the statistical mean of sensitivity and specificity as shown in Equation 5.8 . This metric measures how many samples in both classes are predicted correctly. Similar to sensitivity and specificity, it also ranges from zero to one. The accuracy value of one denotes that the predictor classified all instances of both classes correctly, while, zero accuracy shows that all samples of both classes are misclassified. The mathematical formulation of accuracy is given in equation 5.8.

$$Acc = \frac{TP + TN}{\mathrm{TP+TN+FP+FN}} \tag{4.7}$$

The fourth measure is the Matthews correlation coefficient. It is an ideal metric for data that has imbalance class representations. The values of Matthew correlation coefficient ranges from -1 to +1. The -1 shows that the predictor misclassified all instances of both m6A and non-m6A classes. The Matthews correlation coefficient value of one means predictor classified both sets of samples correctly. The zero Matthews correlation coefficient is nothing but like a random guess.The mathematical description of MCC is given

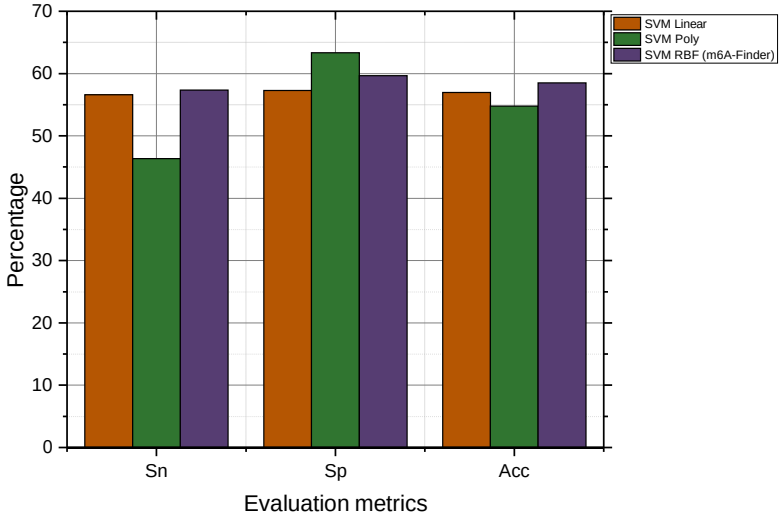


Figure 4.2: Performance of various SVM kernels on all features.

in Equation 5.9.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.8}$$

4.3.3 Performance of *m6A-Finder* on all features

We utilized the support vector machine (SVM) classifier along with varying kernel functions to check their performance on all features. We used three different kernels of SVM i.e., Linear, Polynomial and Radial basis function (RBF) to assess their performance on all features. As shown in Figure 4.2, the RBF kernel showed better results compared to linear and polynomial kernels. These results show that our data has non-linear distribution. Although SVM RBF outperformed other kernels, its results are still low due to overfitting. SVM easily overfits, when features in the dataset are more than the number of samples in the dataset. To solve this problem, we used mRMR to select relevant features and remove redundant and unnecessary features.

4.3.4 Performance of *m6A-Finder* on reduced features

We utilized the mRMR feature selection algorithm to reduce the dimension of data and select optimal features. We utilized different classification algorithms to analyze their

Table 4.3: Comparison of various SVM kernels on all and reduced features.

S#	Features	Classifier	Sn(%)	Sp(%)	Acc(%)	MCC(%)
1	All features	SVM Linear	56.61	57.31	56.96	0.14
2	reduced mRMR features	SVM Linear	79.34	78.27	78.81	0.58
3	All features	SVM Poly	46.36	63.35	54.78	0.10
4	reduced mRMR features	SVM Poly	80.49	81.48	80.99	0.62
5	All features	SVM RBF (m6A-Finder)	57.38	59.68	58.53	0.17
6	reduced mRMR features	SVM RBF (m6A-Finder)	81.02	**82.25**	**81.64**	**0.63**

Table 4.4: Comparison of m6A-Finder with state-of-the-art predictors.

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC(%)
pRNAm_PC[19]	69.72	69.75	69.74	0.40
M6A_HPCS[23]	77.35	67.41	72.38	0.45
RAM-ESVM [24]	78.93	77.78	78.35	0.57
imethyl- STTNC [25]	70.32	68.17	69.84	0.38
iN6-Methyl [58]	77.04	78.50	77.77	0.5550
m6A-Finder	**82.10**	**81.94**	**82.02**	**64**

performance on these reduced features set as shown in Table 4.3. It is clear from Table 4.3, that SVM based on the RBF kernel showed better results than all other classifiers. The reason behind its excellent performance is the low dimension of data in which it draws a clear hyperplane between classes.

4.3.5 Comparison of m6A-Finder with state-of-the-art predictors

We also compare our results with recently developed state of the art predictors for the same problem. We picked the same benchmark dataset and the same validation technique to keep fairness in comparison. We employed four widely used metrics for comparison purposes. The comparative analysis in Table 5.2 shows that our predictor outperforms *pRNAm_PC* [19], *M6A_HPCS* [23], *RAM-ESVM* [24], *imethyl-STTNC* [25] and *iN6-Methyl* [58] in all four performance metrics. The reason behind this outstanding performance is that our features are more discriminative than other features.

4.4 Discussion

N6-methylation (m6A) is one of the most frequent modification among all RNA modifications. It controls many biological processes like messenger-RNA splicing, micro-RNA regulation, cell reprogramming, cell differentiation and cell stability etc. Thus, its correct identification is necessary for in depth understanding of many biological processes.

The existing predictors are facing a number of challenges in the classification of m6A sites. First of all, the structure of yeast transcriptome is complex as compared to other transcriptomes. The second reason for wrong prediction is the existing feature encoding schemes that do not capture hidden-patterns surrounding m6A sites. The third reason is that the state-of-the-art techniques used only short order local features. They do not utilize long order local features with global sequence order features.

We explain via a working example, as shown in Figure 4.3. First of all, it takes an input sequence of RNA of length 51bp and extracts hexa-nucleotide composition features and ACF features based on physical properties. The hexa-nucleotide composition features have long-range local information of the sequence, while ACF features have global long-range sequence order details. By using these two techniques, we acquire both local and global sequence order information, but they also increase the vector dimension. To solve this problem, we utilize the mRMR feature selection technique. The mRMR selects features which have a high correlation with the target class and removes redundant features. Now we have features which can discriminate m6A sites from non-m6A sites. Finally, we give this reduced feature vector to our classifier to predict the presence or absence of m6A site in a given sequence. In the next subsection, we discuss the feature size effect on the performance of the predictor.

4.4.1 Impact of Feature Selection

We use both long-order local features as well as global features. Although the combination of these two set features captures hidden patterns surrounding m6A sites, but also cause overfitting. To select discriminative feature without overfitting, we use the mRMR feature selection algorithm. The mRMR selects features that have high relevance to the

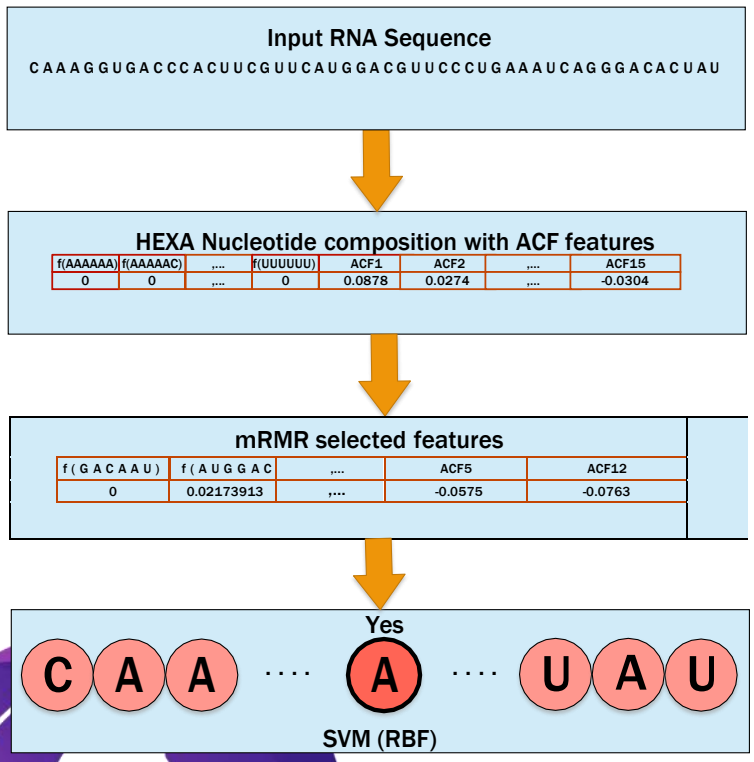


Figure 4.3: Working Example of the m6A-Finder

target class and minimum redundancy with each other. The mRMR takes K as an argument which is the number of features to be selected. The value of K is domain-specific.

To select an optimal value of K, we start from 50 features and increment these features by 50. From 50 to 150, as shown in the Figure 4.4, performance improvement can be observed in all four metrics. The main reason is that all features in this range are more relevant and less redundant. At 200 features sensitivity decreases and specificity increases. At this point, there are some features which increase the false-negative rate. At 250 and 300 features, again performance improvement can be seen in all four metrics. At 300 features predictor is at the peak of its performance. Now we have optimal features which can distinguish between m6A and non-m6A sites. After this point, including more features cause misprediction in both positive and negative class due to some redundant features.

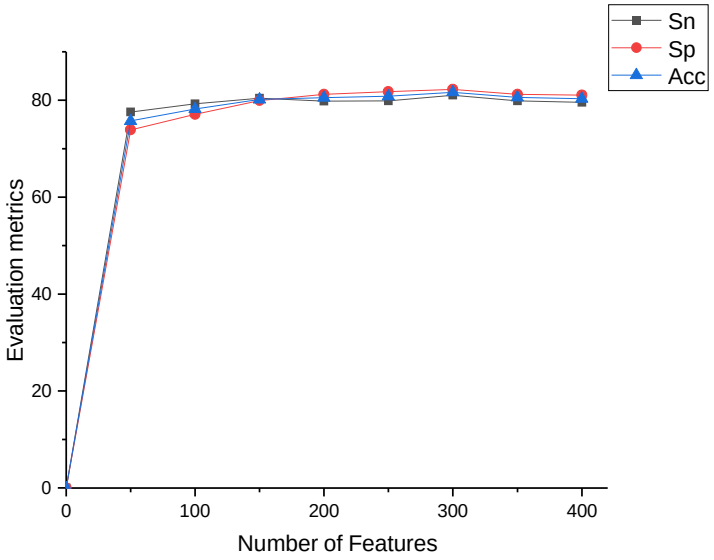


Figure 4.4: Impact of Feature Selection

Table 4.5: Comparison of m6A-Finder with state-of-the-art predictors.

Predictor	Sn(%)	Sp(%)	Acc(%)
iMethyl-RNA[66]	62.80	66.25	67.19
MethyRNA[65]	77.79	100	88.39
m6A-Finder	88.83	89.24	88.81

4.4.2 Performance of the m6A-Finder on Mus musculus data

To examine the generality of the proposed predictor, we validated our predictor on Mus musculus data. The dataset is taken from [65]. The dataset has 725 positive sequences of the m6A sites and 725 sequences of the non-m6A sites of Mus musculus species. As seen in Table 4.5, our predictor outperforms state-of-the-art methods in Mus musculus dataset in the term of accuracy and sensitivity. Our specificity is slightly low due to some features which cause false positive predictions. The performance of our predictor is approximately equal in sensitivity and specificity , which shows that our predictor has patterns that represent both classes equally. These results show that m6A-Finder is a powerful tool and can be used for the identification of m6A sites in other species.

ACADEMIC SOLUTIONS

Chapter 5

m6A-Class: Identification of m6A Sites in Saccharomyces Cerevisiae using Reduced Hybrid Features

5.1 Introduction

The N6-methyladenosine (m6A) modification is frequent modification among all types of cellular RNA modifications [1]. It occurs on position-6 nitrogen in adenosine base and is denoted as m6A in literature. The m6A modifications become swift by various types of catalyzers, namely N6-adenosyl methyltransferases, such as, METTL3, METTL14 and WTAP, etc [2]. Discovered in 1970, m6A sites can be found in both prokaryotes and eukaryotes [1] genomes. These transcription modifications control a lot of gene regulation processes, such as, cell reprogramming and differentiation, primary miRNA regulation, messenger RNA regulation, etc [3]. Any abnormal change in these modifications can cause many diseases, such as, cancer, brain disorder, heart failure, and a lot of other abnormalities [8]. Findings from research suggest that m6A modification occurs near stop codon toward 3' UTR in log internal exon. Therefore, it implies that the m6A modifications are non-randomly distributed and this non-randomness is highly conserved in both prokaryotes and eukaryotes genomes [9]. Thus, the correct identification of m6A sites is vital for understanding the functional mechanisms of both prokaryotes and eukaryotes

genomes.

Recently, wet-lab experiments like m6A-Seq [12] and MeRIF-Seq[8] produced genome-wide m6A modification profiles for humans, Mus musculus, and Saccharomyces cerevisiae (yeast) species. Their results are acceptable for human species, but for yeast and other species, they are not accurate and satisfactory because of complex patterns surrounding m6A sites. They are time-consuming and costly. Therefore, a reliable computational tool is needed for the accurate identification of m6A sites.

To overcome this problem, recently two predictors [13] and [18] are developed to identify m6A sites in Saccharomyces cerevisiae genome. In [13], the authors used nucleotide chemical properties based features to encode m6A signatures. The proposed predictor achieved good specificity but sensitivity was low. In [18] , the authors utilized statistical properties based features using pseudo nucleotide composition (PseNC). The reported sensitivity was good but specificity was low.

Likewise, another predictor pRNAm-PC [19] was developed for the same problem. In the proposed predictor, authors injected chemical properties based values of nucleotides into auto-cross variance and auto-covariance techniques for feature extraction. The mentioned method improved specificity and accuracy, but sensitivity was low compared to previous methods.

Inspired by [13, 18], SRAMP [20] was proposed for classification m6A sites in mammalian data. The authors used a binary encoding scheme to convert 16 dinucleotide pairs. The proposed predictor achieved good specificity but has low sensitivity. Likewise SRAMP, another predictor methyRNA [21] was developed for the same problem. The proposed predictor produces acceptable results on the mammalian dataset, but for yeast data, it yields low accuracy.

For further improvement in this field, a heuristic approach is applied to extract chemical features from the input m6A sequences [23]. The performance of the predictor is evaluated on Saccharomyces cerevisiae data. The predictor sensitivity was somehow favorable, but specificity was low compared to other predictors. To overcome the problem faced by [23], RAM-ESVM [24] was introduced. They used three different types of feature extraction to build individual models based on support vector machines. The mention predictor

slightly improved the performance, but further enhancement is needed.

More recently, methyl-STTNC [25] was proposed to find m6A sites in multiple species. The authors used "split-trinucleotide and split-tetranucleotide" frequency compositions to extract features from input sequences. The proposed method performed well on mammalian data, but its accuracy was very low for Saccharomyces cerevisiae data. Similar to imethyl-STTNC, another method BERMP [26] was proposed to address the same problem. They used Enhanced NucleicAcid Composition (ENAC) for feature extraction. The XGBoost classifier was used for classification. The performance of the predictor is better on the mammalian dataset, its results are poor on the Saccharomyces cerevisiae dataset.

Like imethyl-STTNC [25] and BERMP [26], the M6A-MRFS [14] predictor was introduced to detect m6A sites in cross-species transcriptomes. The Dinucleotide frequency composition based on Position-specific information and binary encoding scheme, were used for feature extraction. The accuracy of this method was low on yeast data. Another predictor named iN6-Methyl [67] was proposed for cross-species data. They used the word2vec technique to extract features from RNA sequences and CNN for classification. The proposed predictor performs well on mammalian datasets, but its performance is low on Saccharomyces cerevisiae dataset due to non-discriminative features. For detecting m6A sites in Saccharomyces cerevisiae, the mentioned feature extraction methods in literature are not well enough to captures hidden insights from input sequences because complex patterns surrounding m6A sites. Moreover, researchers have not studied the surrounding information of the m6A sites in Saccharomyces cerevisiae transcriptome with detail [20].

Recently, iRNA-Freq [68] was introduced to identify m6A in Saccharomyces cerevisiae transcriptome. The frequent gapped-Kmer was used for feature extraction and linear regressor for classification purposes. The predictor has acceptable accuracy, but sensitivity is very low compared to previous predictors. Most recently, another group of researchers proposed iMRM [69] to find different RNA modification sites in various species. They used six different types of statistical and chemical properties based techniques for feature extraction. They used an incremental feature selection technique to select optimal features. Finally, XGboost is used as a classifier. The proposed predictor has performed

well on all data except Saccharomyces cerevisiae. The feature extraction techniques used in this method are not enough to discriminate m6A sites from non-m6A sites in yeast because of complex patterns surrounding m6A sites.

From the detailed review, we conclude that feature extractions have an important role in predictor's performance. The feature representation methods addressed in literature, are not sufficient to identify m6A sites in saccharomyces cerevisiae transcriptome. Besides, statistical and chemical-based features, physical property based features are not used for backer's yeast data. In this paper, we propose a novel predictor named m6A-Class. The proposed predictor employs the fusion of statistical, chemical and physical property based features to detect m6A sites correctly. The proposed predictor solves problems faced by existing predictors due to non-discriminative features. Although, these features capture useful hidden details surrounding m6A sites, but also leads to high dimension feature vector, which results in overfitting. To deal with overfitting, we use a feature selection algorithm to select relevant features and throw away redundant ones.

The key contributions of this study are presented bellow:

1. Novel physical properties based features are combined with statistical, chemical features to capture useful patterns surrounding m6A sites.
2. Filter based feature selection approach, i.e., Minimum Redundancy Maximum Relevance (mRMR) is utilized to include relevant features and discard redundant ones.

The rest of the paper is structured as follows: we explain the materials and methods of the *m6A-Class* in Section 2. In Section 3, we present our results on RNA transcript of *Saccharomyces cerevisiae* species. The Section 4 has detail discussion. Finally, we conclude our study in Section 5.

5.2 Materials and Methods

In this section, we discuss the benchmark dataset as well as the complete architecture of the proposed predictor. In sub sections, we discuss feature extraction, feature selection and classification in detail. In feature extraction step, we explain feature extraction

techniques based on physical,chemical and statistical aspects of nucleotides. In a feature selection section, we discuses Minimum Redundancy Maximum Relevance (mRMR) in detail. In classification section, there is brief introduction about SVM.

5.2.1 Benchmark Dataset

The dataset used in this research is obtained from [13]. The dataset has a total of 2614 samples of saccharomyces cerevisiae species. The first half of the dataset is compose of experimentally confirmed m6A sites, while the second half is experimentally confirmed non-m6A sites. These negative samples are taken from 33,280 samples in a random manner to avoid imbalance bias in training set. All samples have an equal length of 51bp and having a similarity of less than 85 percent.

5.2.2 The *m6A-Class* Model

Existing predictors are facing the problems in the detection m6A sites in Saccharomyces cerevisiae species because of complex patterns surrounding m6A sites and non-discriminative feature representation techniques to capture these patterns. To tackle these problems, we propose a novel computational tool to classify m6A sites. The novelty of the proposed method is that it utilizes physical properties for feature extraction with combination of statistical features and chemical properties based features. The proposed predictor has three key steps i.e, feature extraction, feature selection and classification. In feature extraction, we calculate physical properties based features, statistical compositions features and chemical properties based features. Although, features fusion capture complex patterns surrounding m6A sites, but also increases vector dimension, which causes overfitting. We utilize the feature selection algorithm to solve this problem. Finally, SVM predicts whether RNA sequence has m6A site or not based on these optimal features. The graphical representation of the model is given in Figure 5.1. The key steps of the *m6A-Class* are presented bellow:

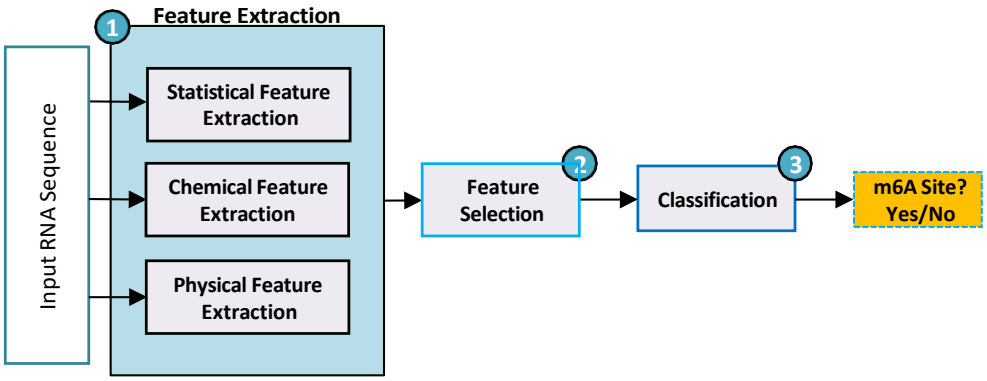


Figure 5.1: The diagrammatic representation of the m6A-Class for prediction of m6A sites in Saccharomyces cerevisiae transcriptomes.

5.2.2.1 Feature extraction

In this step, significant insights are derived from data. First of all, the input sequence of m6A site encoded into numerical values. Then feature are extracted on base of these numerical values and these features make a feature vector. The total number of features in a vector is known as vector dimension. In the area of computational biology, Omic samples are encoded and then converted to numerical values, which capture the hidden information inside samples. In the current problem of the m6A site classification, we calculate hybrid features based on *physical characteristics*, and *statistical composition* and chemical properties to capture surrounding information of m6A site. The fusion of these diverse features, represent encoded information sounding m6A site in a precise way. The brief introduction of these feature representation techniques is provided below:

Statistical Features: Inspired by its excellent performance of statistical composition features in 4mC sites prediction [70], we incorporated novel hepta-nucleotide composition features for input sequences. In this feature extraction method, seven adjacent nucleotides are counted to a single entity and then its frequencies are calculated. As we know, a m6A site sequence has four distinct bases. Therefore, there will 4^7=16,384 features. The mathematical representation of this scheme as follows:

$$R = [f(AAAAAAA),\ f(AAAAAAC),\ f(AAAAAAG),\ f(AAAAAAU), \ldots,\ f(UUUUUUU)]^T \tag{5.1}$$

Where R is feature vector, *f(AAAAAAA)* is the frequency of seven adjacent AAAAAAA nucleotide pair, *f(AAAAAAC)* is the frequency of seven adjacent AAAAAAC pair and so on.

Chemical Propertes based Features: Encouraged of its good results in other modifications sites, we also extracted features based on the chemical property (also called Ring function hydrogen chemical property) [36]-[37]. In the mentioned scheme, a pair of nucleotides is associated with groups, and these groups are encoded into binary values. If we consider adenine and guanine they come under purines group because of double rings. On the other hand, cytosine and uracil come under pyrimidines group as they have a single ring. Another property for grouping is a secondary structure in which hydrogen bond is considered to make groups. Hence, cytosine and guanine have a strong bond. Therefore, they are a member of a strong hydrogen bond group. While the remaining two belongs to the weak hydrogen bond group. If we consider chemical functionality, the adenine and cytosine come under the amino group. On the hand, guanine and uracil are considered as keto.

To encode the m6A Sequence based on these chemical properties, we make use of three coordinate systems, i.e., x, y and z. We use x coordinate for ring structure, y coordinate for hydrogen bond, and z coordinate for chemical grouping. The three coordinates system can be formulated as follow:

$$xi = \begin{cases} 1 & if\ si \in \{A, G\} \\ 0 & if\ si \in \{C, U\} \end{cases}$$

$$yi = \begin{cases} 1 & if\ si \in \{A, U\} \\ 0 & if\ si \in \{C, G\} \end{cases}$$

$$zi = \begin{cases} 1 & if\ si \in \{A, C\} \\ 0 & if\ si \in \{G, U\} \end{cases}$$

As an illustration, A is converted to (1, 1, 1), C is converted to (0, 1, 0), G is converted

Table 5.1: List of physical properties associated with each nucleotide and their corresponding values.

Physical property	A	C	G	U
Density	2.1	1.6	2.2	1.5
Boiling point	676.3	445.8	561.5	440.5
Melting point	234	325	360	330
Flash point	362.8	223.4	293.4	220.02
Mass	267	111.04	151.04	112.02
PSA	139.54	71.77	100.45	65.72
logP	-1.02	-2.29	-2.03	-2.55
Index of Refraction	1.907	1.689	2.047	1.64

to (1, 1, 1), and U is converted to (0, 0, 0). Besides chemical properties based features, we also include nucleotide density composition features [39]. It is denoted by d_i. To calculate d_i of nucleotide q at position i, the following formula is used:

$$di = \frac{1}{|Ni|}\sum_{n=1}^{|Ni|} f_n(q_i);\ where\ f_n(q_i) = \begin{cases} 1; & if\ q_i\ is\ present\ at\ location\ n \\ 0; & otherwise. \end{cases} \quad (5.2)$$

Where, $|Ni|$ is the length of the *ith* prefix substring $n_1, n_2, .., n_i$ in the input m6A RNA sequence, and q_i may be any nucleotide at *ith* position and $q_i \in$(A, C, G, U), while $f_n(q_i)$ is total number occurrences of a nucleotide q_i at location *n*. Hence every letter is represented by four values, so finally we will get 4× 51=204 chemical properties based features.

Physical Feature Extraction: The hepta-nucleotide composition features and chemical properties base features only capture local order details of sequence, they ignore global order information. To capture global details, we extract physical property based features. Therefore, we pick out eight physical properties of individual nucleotides i.e., density, boiling point, melting point, flash point, mass, Polar surface Area (PSA), Partition coefficient (logP), and index of refraction. The detail about these properties is given in Table 5.1.

We constructed feature vector by modeling variance between physical properties in the given sequence over lagged version by itself. To do so, we use autocovariance function

(acovf) to encode physical property values into a feature vector, as shown in Equation 5.3.

$$acovf(P_j, k) = \frac{1}{(n-k)} \sum_{i=1}^{n-k} (R_i^{P_j} - \mu)(R_{i+1}^{Pj} - \mu) \tag{5.3}$$

where, P_j is jth physical property value with $j \in [1, 2, 3, \ldots, 8]$, k is the lag value for auto-covariance, with $k \in [1, 2, 3, \ldots, 50]$, n is the length of the input RNA sequence, μ is the mean value of the input sequence and R_i^{Pj} is the value of residue R_i with property P_j.

In order to keep values separated for each lag, the values obtained for each property from Equation 5.3 are averaged for each lag using Equation 5.4.

$$F_{physical}(k) = \frac{1}{Z} \sum_{j=1}^{Z} ACF(P_j, k) \tag{5.4}$$

where, Z is total physical properties, in our case Z=8 and $F_{physical}(k)$ is a feature value for each lag k. This step calculates 50 features, we selected 50 because of including all lags values of sequence. After this step, we combine physical properties based values with chemical and statistical features, which result in feature vector having dimension of 16,638 features. Although these features capture useful hidden information surrounding m6A sites, but also increases vector dimension, which causes overfitting. To resolve this issue, we use feature selection algorithm to include relevant features and discard redundant ones.

5.2.2.2 Feature selection

In this work, different feature extraction schemes are used to extract feature from the RNA sample, which leads to a high dimensional vector. To solve this problem, we used a feature selection technique to reduce dimension as well as reduce the computational burden. There are two major categories of techniques for feature selection: filter method and wrapper methods. The first one calculates an internal correlation between features as an evaluation measure. The second one takes a subset of features to check its performance on the classifier. The filter method has low time complexity as compared to wrapper methods. Another problem of wrapper methods is that sometimes they suffer from over-

fitting. Due to these benefits of filter based techniques, we utilized Minimum redundancy maximum relevance feature (mRMR) to select relevant feature.

The mRMR is a filter method based technique for feature selection, which was developed by [71]. The basic mechanism of this technique is to rank features which have high relevance to target class. For example at iteration i, the score-$FCQ_i(f)$ for each feature is calculated by following formula:

$$score\_FCQ(f) = \frac{F(f|tc)}{\sum_{s \in featuresselected\ until\ i-1} |cor(f,s)|/(i-1)} \tag{5.5}$$

The feature that has highest score among all features at iteration i is selected, while score-$FCQ_i(f)$ is F-test with a correlation quotient. In Equation 5.5, F is F-statistic used to calculate relevance between feature f and target class i.e., tc, while redundancy between feature f and already selected features is calculated using function *cor*. The function *cor* is actually Pearson correlation.

For the current problem, the mRMR selects 400 from from fusion of three types of features. Finally, these 400 features go to SVM to predict whether the center adenosine in input sequence is methylated or not.

5.2.2.3 Classification

We use a Support vector machine (SVM) as a classifier to predict m6A sites and non-m6A sites in Saccharomyces cerevisiae transcriptome[61]. It draws a hyperplane between classes based on statistical theory. The SVM maximizes marginal distance from hyperplane to improve prediction [62]. It consists of different kernel functions, which can be utilized according to the distribution of data [63]. For the current problem, we use three kernels (linear, polynomial and radial basis function kernel). The radial basis function kernel (RBF) outperforms other kernels that show our data has non-linear distribution. For the current problem, The SVM (RBF) takes 400 features as input and draws a hyperplane according to its kernel function. Finally, the presence or absence of the m6A site is predicted on the basis of these features.

5.3 Experimental Setup and Results

In this section, we explain the validation method utilized for this study as well as we discuss different metrics used for comparison. In further subsections, we analyze the performance of m6A-Class on combined and reduced features. Finally, we compare *m6A-Class* with state-of-the-art predictors.

5.3.1 Jacknife validation

The power of predictor is judged upon its prediction's quality. Therefore, it is necessary to evaluate the performance of the predictor carefully. There are three types of methods for performance evaluation, which are hold-out, cross-validation and Jackknife. The jackknife technique is famous amongst them due to low bias in results. In the current study, we also pick the jackknife validation method for performance evaluation. In this method, at a time, only one instance is chosen for testing while others remained for training. There are total N folds, where N is total instances in the dataset. Finally, all predicted labels are combined to get final results.

5.3.2 Performance Metrics

Our problem is supervised binary classification task, therefore, m6A-Class will predict the presence or absence of m6A site in a given sequence. There are a lot of evaluation metrics which can be used to measure the predictor performance. The sensitivity, specificity, Accuracy, and Mathew correlation coefficient are commonly used amongst them. For this problem, we also use these four metrics. These metrics are based upon four values of the confusion matrix, which are true-positive, true-negative, false-positive, and false-negative. These values can be represented by TP, TN, FP and FN respectively [72]. The mathematical representations of these four metrics are discussed below.

The sensitivity metric measures how many positive samples are correctly classified among all samples of positive class. The sensitivity metric has lowest value zero and the highest

is one. When sensitivity=1 means all samples in positive class are identified as m6A site. When sensitivity=0 means none of the m6A sites is identified as m6A site. The sensitivity metric eludes false negative as show in Equation 5.6.

$$Sn = \frac{TP}{\text{TP + FN}} \tag{5.6}$$

The specificity metric measures how many negative samples are correctly classified among all samples of negative class. Similarly to sensitivity, specificity metric has lowest value zero and the highest is one. When specificity=1 means all samples in the negative class are identified as non-m6A sites. When Specificity=0 tells that none of the non-m6A are identified correctly. Unlike sensitivity, specificity eludes false positive as shown in Equation 5.7.

$$Sp = \frac{TN}{\text{TN + FP}} \tag{5.7}$$

Accuracy in a simple word is an average of sensitivity and specificity as shown in Equation 5.8. Likewise, sensitivity and specificity, Accuracy also range from zero to 1. The accuracy=0 indicates that none of the instances in both classes are identified correctly. The accuracy=1 reports that sequences in both classes are identified accurately.

$$Acc = \frac{TP + TN}{\text{TP+TN+FP+FN}} \tag{5.8}$$

Likewise accuracy, another essential metric to check the predictor's quality is Mathew correlation coefficient (MCC). It is well suitable when data has imbalance classes. Unlike, three other measures it ranges from -1 to +1. The MCC=-1 points out that all samples in both groups are classified incorrectly. When it has a value equal to +1 shows that all instances of both classes are classified correctly. If MCC has a value equal to zero , it is like a random prediction. Its mathematical formulation is shown in Equation 5.9 .

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.9}$$

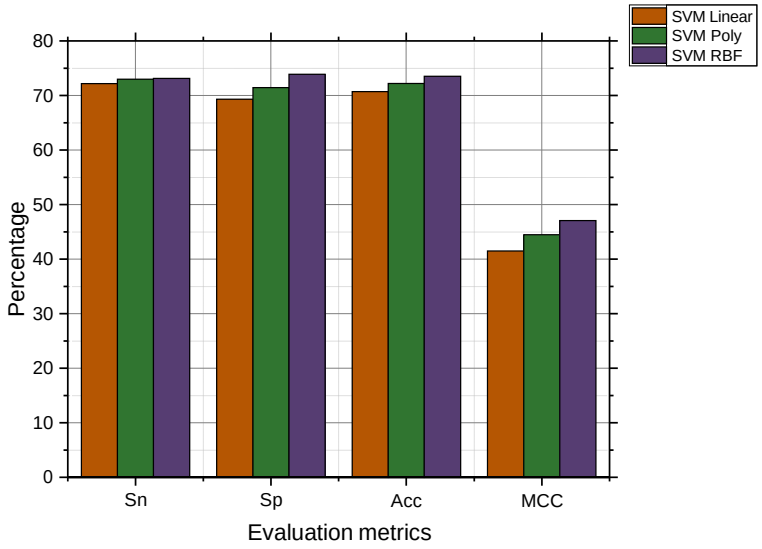


Figure 5.2: Performance of *m6A-Class* on on fusion of features for various SVM kernels.

5.3.3 Performance of *m6A-Class* on fusion of features using 5 cross validation

We used different kernels of SVM to analyse its performance on all features. We selected three different kernel functions of SVM i.e., Polynomial, Radial basis function (RBF) and Linear. Due to the non-linear distribution of data, the RBF kernel outperformed the other two kernels, as shown in Figure 5.2. Although SVM based on RBF kernel produced better results than Linear and Polynomial kernels, its results are still low because of overfitting. The SVM easily suffers from overfitting compared to tree-based classifiers, when data has a high dimension. To resolve this issue, we use mRMR to include relevant features and discard redundant ones.

5.3.4 Performance of our predictor on optimal features using 5 cross validation

As discussed in section 2.2, we reduced features dimension via feature selection algorithm. For that purpose, we use famous mRMR algorithm. We employed different kernels of SVM to check their performance on these reduced features. The results in Figure 5.3

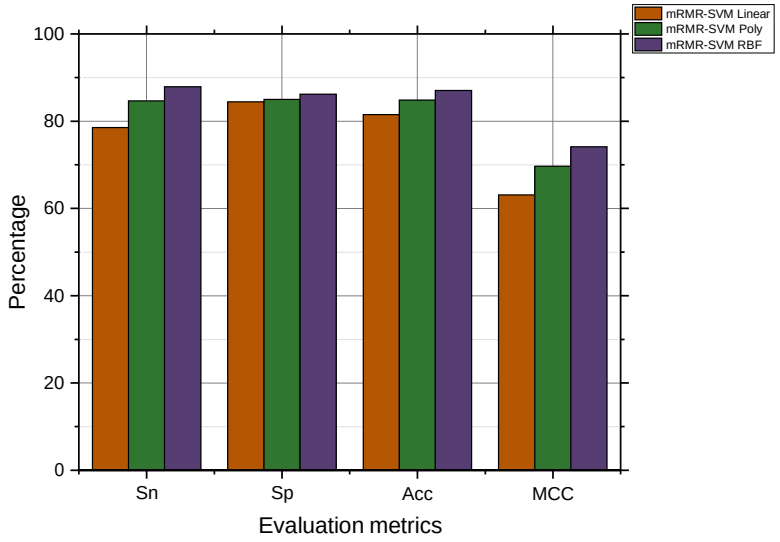


Figure 5.3: Performance of *m6A-Class* on optimal feature types for various SVM kernels.

show that SVM based on Radial basis function kernel performs better compared to other kernels. If we match the results in Figure 5.2 and Figure 5.3, it is clear that feature selection algorithm increases the performance of the predictor. Therefore, it is obvious that feature selection has an indispensable role in performance improvement. Besides this key benefit, it also reduces time complexity.

5.3.5 Comparison of m6A-Class with state-of-the-art predictors.

In current section, we compare m6A-Class with existing predictors which produced acceptable results for mentioned problem. It can be seen in Table 5.2, For comparison purpose, we select four widely used metrics. We compared newly developed predictor to state-of-the-art predictors i.e. pRNAm_PC [19], M6A_HPCS [23] , RAM-ESVM[24], imethyl-STTNC [25], M6AMRFS [14], iN6-Methyl[67] and iMRM [69] .

To maintain fairness in comparison, same dataset and same cross validation (i.e., Jackknife) are utilized. As shown in Table 5.2, the *m6A-Class* predictor outperform all of the recently developed tools. The *m6A-Class* is superior in performance from pRNAm_PC [19], M6A_HPCS [23] , RAM-ESVM[24], imethyl-STTNC [25], M6AMRFS [14], iN6-Methyl[67] and iMRM [69] in all four performance metrics.

Table 5.2: Comparison of m6A-Class with state-of-the-art predictors.

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC
pRNAm_PC[19]	69.72	69.75	69.74	0.40
M6A_HPCS[23]	77.35	67.41	72.38	0.45
RAM-ESVM[24]	78.93	77.78	78.35	0.57
imethyl- STTNC[25]	70.32	68.17	69.84	0.38
M6AMRFS[14]	75.21	73.39	74.25	0.4852
iN6-Methyl[67]	77.04	78.50	77.77	0.5550
iMRM [69]	76.15	74.62	75.38	0.5078
m6A-Class	**88.45**	**86.30**	**87.38**	**0.7477**

5.4 Discussion

N6-methyladenosine (m6A) is one of the most frequent post-transcriptional modification. It occurs in both eukaryote and prokaryote mRNA transcripts. These modifications regulate messenger RNA because RNA binding proteins recognize these specific modifications. These modification sites perform important roles in different biological processes,like, microRNA regulation, messager RNA splicing, cell stability, cell differentiation and cell reprogramming, etc. For in depth understanding of these biological processes, the correct identification is necessary.

The current state-of-the-art predictors are facing problems in detection of m6A sites in yeast transcriptome. First reason of wrong predictions is that there are complex patterns around m6A sites in yeast species compared to other species. Another reason for false predictions is that, they either utilized chemical or statistical features. They do not use fusion of features to capture useful insights. The third reason is that the existing predictors neither utilized long-range local sequence features in solo nor with fusion with global features.

How our predictor solves these problems, we demonstrate it through working example, as shown in Figure 5.4. In the first step, we give an RNA sequence having a length of 51 to the predictor. The predictor extracts statistical, chemical and physical features from a given sequence. As we known that, the fusion of these three type of features captures hidden useful details surrounding m6A sites, but also results in high dimension vector, which causes overfitting in our case. In third step, we solve this problem by

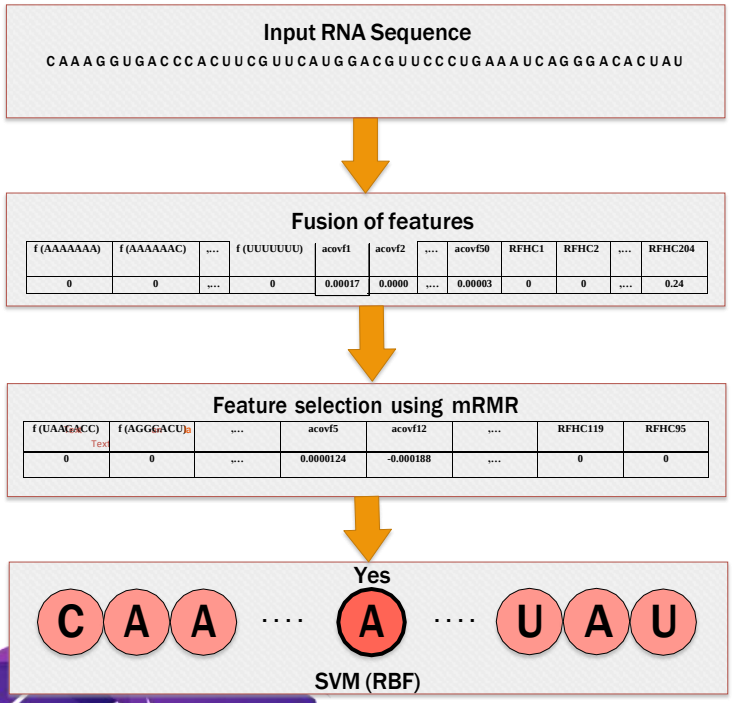


Figure 5.4: Working Example of m6A-Class to detect a m6A site.

using feature selection algorithm i.e, mRMR. The mMRMR feature selection algorithm selects a feature at every iteration, which have high relevance to target class and minimum redundancy to already selected features in previous iterations. After this step, we have 400 features which are more relevant to target class and less redundant to each other. Finally our predictor uses SVM to predict whether the input sequence contains m6A site or not on the basis of these features. In the next subsection, we analysis the performance of predictor on different size of features.

5.4.1 Enhancements in finding m6A sites

We extract different types of features from the input RNA sequence. Although this diverse combination of features capture useful patterns surrounding m6A sites but also results in overfitting, to solve this problem, we use mRMR. The mRMR selects relevant features and discards redundant ones. How many features will be selected by mRMR, it is totally domain-specific.

For the current problem, we start from 50 features and increment by 50. If we compare the

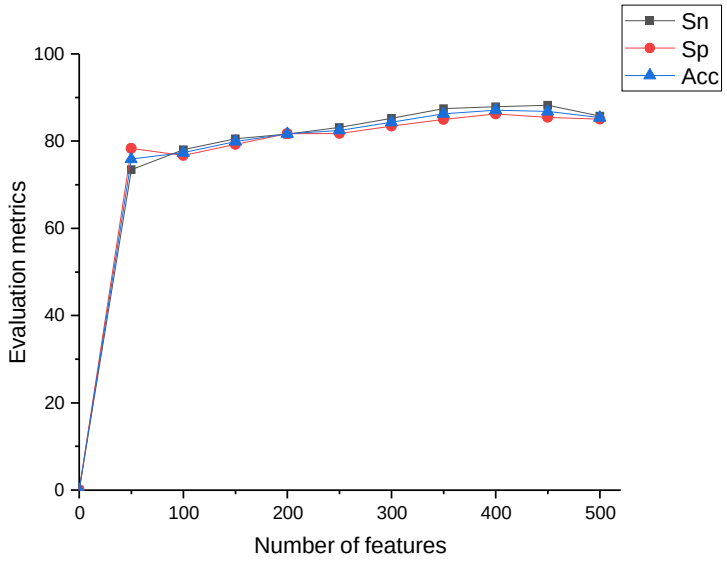


Figure 5.5: Enhancements in finding m6A sites

performance of the classifier on 50 and 100 features, accuracy and sensitivity are increasing, but specificity is decreasing. The degradation in specificity shows that some features increase the false-positive rate. From 150 to 400 features, performance improvement can be seen in all metrics. At 400 features, the predictor achieves the highest accuracy. Now we have features that are more discriminative. After this point, adding more features result in performance degradation as shown in 5.5.

5.4.2 Insights for the detection of m6A

To find insights inside m6A and non-m6A sequences, we have analyzed them based on statistical, chemical and physical properties. To observes patterns based on statistical features, we have checked the same type of Condons left and right sides of the central GAC motif. We have observed that AAG Condon has a 46% chance of occurrence in m6A-sequences and 11% in non-m6A sequences. Similarly, the CAA, UAA and UUG have 11% frequency in non-m6A sequences and a 0% frequency in m6A-sequences.

To analyze the sequences on chemical properties, we have observed the same pairs of nucleotides on both sides of the central GAC motif. The results show that the amino group has the most occurrences in non-m6A sequences than non-m6A sequences.

Similarly, to check insights base on physical properties, we have analyzed the same type of nucleotides of both sides of the central GAC motif. The results show that non-m6A sequences have a greater number of lightweights nucleotides than m6A-sequences.

Chapter 6

Conclusion and Future Work

The N6-methyladenosine (m6A) is the most abundant RNA modification. It performs a vital role in different biological processes like messenger RNA translation, micro RNA regulation, cell stability, cell reprogramming,cell differentiation and X-chromosome inactivation etc. Any abnormal mutation in the m6A site can cause heart attack, obesity, cancer and brain tumor, etc. Therefore, the correct detection of m6A sites in yeast transcriptome is very necessary to overcome these diseases. The existing feature extraction techniques are not enough to discriminate m6A sites from non-m6A site. They still have gap for further improvement. In this chapter, we conclude our study by summarizing the main contributions and also discuss possible future extensions of this work for further improvements.

6.1 Conclusion

The existing predictor utilize features which are not sufficient to distinguish m6A sites from non m6A sites in yeast species The main reason for false prediction is that, there are complex patterns surrounding m6A sites in yeast species compared to other species. The second reason for high false positive and false negative is that, they either utilize statistical or chemical features. The fusion of these two types of features is not widely explored to capture hidden insights surrounding m6A sites. The third and most import reason for

false prediction rate, they are using short-range local sequence order information. They neither use long-range local sequence order information independently nor with fusion global features based physical properties.

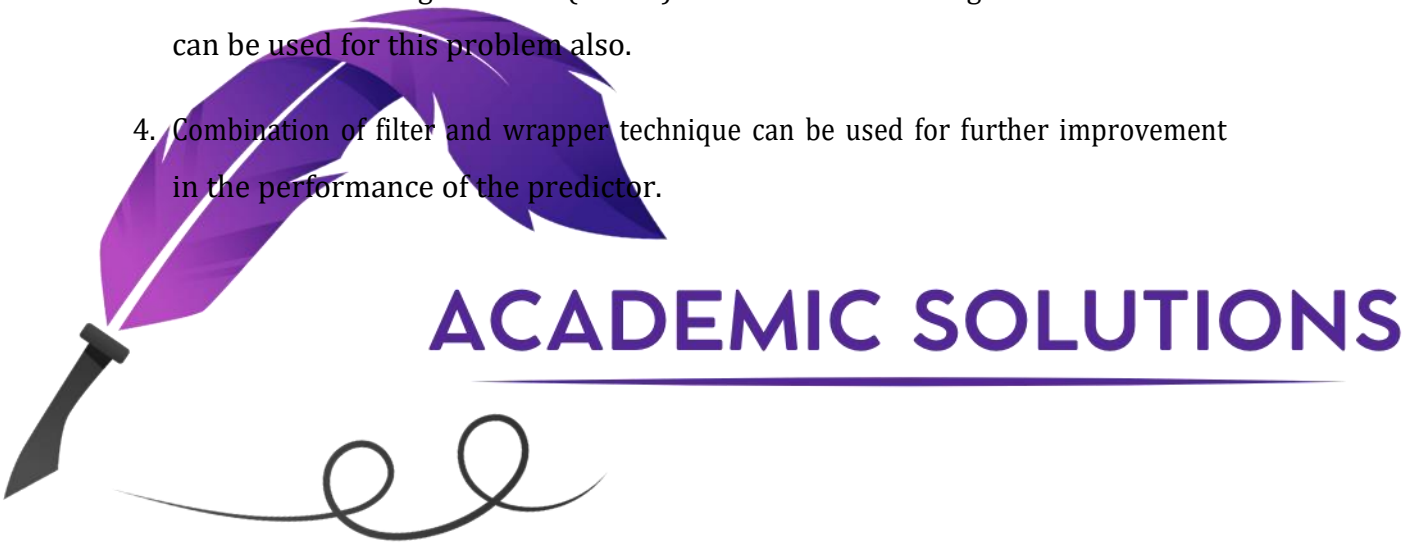
In this work, we propose three predictors for the classification m6A site and non-m6A sites. The first one is named m6A-Pred, and the second and third are named m6A-Finder and m6A-Class, respectively. The m6A-Pred uses the fusion of statistical (di-nucleotide composition and tri-nucleotide composition) and chemical properties based features (ring function hydrogen chemical property). As we know, fusion increases the feature vector dimension, which causes overfitting for current problem. To solve this problem, we use the Genetic algorithm (GA) as a feature selection algorithm. The genetic algorithm selects optimal features and discards unnecessary ones. Finally, we use the Random forest for prediction. In m6A-Finder, we extract features based on statistical properties (Hexa-nucleotide composition) and physical properties (Autocorrelation features based on physical properties). Although, hexa-nucleotide composition features capture long-range local sequence order information, but also increases feature vector dimension, which leads to overfitting in current scenario. To solve this problem, we use Minimum redundancy maximum relevance (mRMR) to select optimal features and remove unnecessary and redundant features. Finally, we use the Support vector machine to predict whether the input RNA sequence has an m6A site or not.

All of the proposed predictors outperform existing state-of-the-art predictors. The m6A-Pred showed better sensitivity, accuracy and MCC compared to existing techniques and the m6A-Finder outperforms in all four metrics. This performance suggests that our feature extraction techniques are more discriminative than existing approaches. Therefore, we cay say our proposed predictors are more accurate in finding m6A sites in yeast. Besides outstanding performance, our predictors have low time complexity due to feature selection. Finally, it is concluded that our predictors can be used as useful biomedical tools in basic research and drug discovery.

6.2 Future Work

The proposed research opens up numerous new directions for future research. Some key points are highlighted below:

1. Long-order local sequential features can be used with existing features techniques for further improvement in accuracy.
2. Deep learning based techniques have shown outstanding performance in other fields and can be used for the same problem.
3. The Continuous Bag of Words (CBOW) has shown outstanding results in NLP and can be used for this problem also.
4. Combination of filter and wrapper technique can be used for further improvement in the performance of the predictor.



ACADEMIC SOLUTIONS

Bibliography

[1] W. A. Cantara, P. F. Crain, J. Rozenski, J. A. McCloskey, K. A. Harris, X. Zhang, F. A. Vendeix, D. Fabris, and P. F. Agris, "The rna modification database, rnamdb: 2011 update," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D195–D201, 2010.

[2] J. Liu and G. Jia, "Methylation modifications in eukaryotic messenger rna," *Journal of Genetics and Genomics*, vol. 41, no. 1, pp. 21–33, 2014.

[3] K. D. Meyer and S. R. Jaffrey, "The dynamic epitranscriptome: N 6-methyladenosine and gene expression control," *Nature reviews Molecular cell biology*, vol. 15, no. 5, p. 313, 2014.

[4] A. Akilzhanova, Z. Nurkina, K. Momynaliev, E. Ramanculov, Z. Zhumadilov, T. Rakhypbekov, N. Hayashida, M. Nakashima, and N. Takamura, "Genetic profile and determinants of homocysteine levels in kazakhstan patients with breast cancer," *Anticancer research*, vol. 33, no. 9, pp. 4049–4059, 2013.

[5] R. Casalegno-Garduno, A. Schmitt, X. Wang, X. Xu, and M. Schmitt, "Wilms' tumor 1 as a novel target for immunotherapy of leukemia," in *Transplantation proceedings*, vol. 42, no. 8. Elsevier, 2010, pp. 3309–3311.

[6] M. J. Machiela, S. Lindström, N. E. Allen, C. A. Haiman, D. Albanes, A. Barricarte, S. I. Berndt, H. B. Bueno-de Mesquita, S. Chanock, J. M. Gaziano *et al.*, "Association of type 2 diabetes susceptibility variants with advanced prostate cancer risk in the breast and prostate cancer cohort consortium," *American journal of epidemiology*, vol. 176, no. 12, pp. 1121–1129, 2012.

[7] K.-J. Heiliger, J. Hess, D. Vitagliano, P. Salerno, H. Braselmann, G. Salvatore,

C. Ugolini, I. Summerer, T. Bogdanova, K. Unger *et al.*, "Novel candidate genes of thyroid tumourigenesis identified in trk-t1 transgenic mice," *Endocrine Related Cancer*, vol. 19, no. 3, p. 409, 2012.

[8] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mrna methylation reveals enrichment in 3' utrs and near stop codons," *Cell*, vol. 149, no. 7, pp. 1635–1646, 2012.

[9] S. Schwartz, S. D. Agarwala, M. R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T. S. Mikkelsen, R. Satija, G. Ruvkun *et al.*, "High-resolution mapping reveals a conserved, widespread, dynamic mrna methylation program in yeast meiosis," *Cell*, vol. 155, no. 6, pp. 1409–1421, 2013.

[10] B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey, "Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome," *Nature methods*, vol. 12, no. 8, p. 767, 2015.

[11] G.-Z. Luo, A. MacQueen, G. Zheng, H. Duan, L. C. Dore, Z. Lu, J. Liu, K. Chen, G. Jia, J. Bergelson *et al.*, "Unique features of the m 6 a methylome in arabidopsis thaliana," *Nature communications*, vol. 5, p. 5630, 2014.

[12] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec *et al.*, "Topology of the human and mouse m 6 a rna methylomes revealed by m 6 a-seq," *Nature*, vol. 485, no. 7397, p. 201, 2012.

[13] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "irna-methyl: Identifying n6-methyladenosine sites using pseudo nucleotide composition," *Analytical biochemistry*, vol. 490, pp. 26–33, 2015.

[14] X. Qiang, H. Chen, X. Ye, R. Su, and L. Wei, "M6amrfs: robust prediction of n6-methyladenosine sites with sequence-based features in multiple species," *Frontiers in genetics*, vol. 9, p. 495, 2018.

[15] 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 *et al.*, "Alkbh5 is a mammalian rna demethylase

that impacts rna metabolism and mouse fertility," *Molecular cell*, vol. 49, no. 1, pp. 18–29, 2013.

[16] G. Keith, "Mobilities of modified ribonucleotides on two-dimensional cellulose thin-layer chromatography," *Biochimie*, vol. 77, no. 1-2, pp. 142–144, 1995.

[17] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec *et al.*, "Topology of the human and mouse m 6 a rna methylomes revealed by m 6 a-seq," *Nature*, vol. 485, no. 7397, pp. 201–206, 2012.

[18] W. Chen, H. Tran, Z. Liang, H. Lin, and L. Zhang, "Identification and analysis of the n 6-methyladenosine in the saccharomyces cerevisiae transcriptome," *Scientific reports*, vol. 5, p. 13859, 2015.

[19] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, and K.-C. Chou, "prnam-pc: Predicting n6-methyladenosine sites in rna sequences via physical–chemical properties," *Analytical biochemistry*, vol. 497, pp. 60–67, 2016.

[20] Y. Zhou, P. Zeng, Y.-H. Li, Z. Zhang, and Q. Cui, "Sramp: prediction of mammalian n6-methyladenosine (m6a) sites based on sequence-derived features," *Nucleic acids research*, vol. 44, no. 10, pp. e91–e91, 2016.

[21] W. Chen, H. Tang, and H. Lin, "Methyrna: a web server for identification of n6-methyladenosine sites," *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 3, pp. 683–687, 2017.

[22] Z. Zhao, H. Peng, C. Lan, Y. Zheng, L. Fang, and J. Li, "Imbalance learning for the prediction of n 6-methylation sites in mrnas," *BMC genomics*, vol. 19, no. 1, pp. 1–10, 2018.

[23] M. Zhang, J.-W. Sun, Z. Liu, M.-W. Ren, H.-B. Shen, and D.-J. Yu, "Improving n6-methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties," *Analytical biochemistry*, vol. 508, pp. 104–113, 2016.

[24] W. Chen, P. Xing, and Q. Zou, "Detecting n 6-methyladenosine sites from rna transcriptomes using ensemble support vector machines," *Scientific reports*, vol. 7, p. 40242, 2017.

[25] S. Akbar and M. Hayat, "imethyl-sttnc: Identification of n6-methyladenosine sites by extending the idea of saac into chou's pseaac to formulate rna sequences," *Journal of theoretical biology*, vol. 455, pp. 205–211, 2018.

[26] Y. Huang, N. He, Y. Chen, Z. Chen, and L. Li, "Bermp: a cross-species classifier for predicting m6a sites by integrating a deep learning algorithm and a random forest approach," *International journal of biological sciences*, vol. 14, no. 12, p. 1669, 2018.

[27] Y. Zhang and M. Hamada, "Deepm6aseq: prediction and characterization of m6a-containing sequences using deep learning," *BMC bioinformatics*, vol. 19, no. 19, p. 524, 2018.

[28] L. Wenzhong, "Sicm6a: Identifying m6a site across species by transposed gru network," *bioRxiv*, p. 694158, 2019.

[29] Y. Zhuang, H. Liu, X. Song, Y. Ju, and H. Peng, "A linear regression predictor for identifying n6-methyladenosine sites using frequent gapped k-mer pattern," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 673–680, 2019.

[30] K. Liu and W. Chen, "imrm: a platform for simultaneously identifying multiple kinds of rna modifications," *Bioinformatics*, 2020.

[31] Y. Yang, P. J. Hsu, Y.-S. Chen, and Y.-G. Yang, "Dynamic transcriptomic m 6 a decoration: writers, erasers, readers and functions in rna metabolism," *Cell research*, vol. 28, no. 6, pp. 616–624, 2018.

[32] T. W. Nilsen, "Internal mrna methylation finally finds functions," *Science*, vol. 343, no. 6176, pp. 1207–1208, 2014.

[33] K. Chen, Z. Wei, Q. Zhang, X. Wu, R. Rong, Z. Lu, J. Su, J. P. de Magalhaes, D. J. Rigden, and J. Meng, "Whistle: a high-accuracy map of the human n 6-

methyladenosine (m6a) epitranscriptome predicted using a machine learning approach," *Nucleic acids research*, vol. 47, no. 7, pp. e41–e41, 2019.

[34] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A. Dehzangi, "Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences," *Bioinformatics*, vol. 35, no. 19, pp. 3831–3833, 2019.

[35] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb *et al.*, "ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data," *Briefings in bioinformatics*, vol. 21, no. 3, pp. 1047–1057, 2020.

[36] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "irna-ai: identifying the adenosine to inosine editing sites in rna sequences," *Oncotarget*, vol. 8, no. 3, p. 4208, 2017.

[37] W. Chen, P. Feng, H. Tang, H. Ding, and H. Lin, "Rampred: Identifying the n 1-methyladenosine sites in eukaryotic transcriptomes," *Scientific reports*, vol. 6, p. 31080, 2016.

[38] A. G. Bari, M. R. Reaz, H.-J. Choi, and B.-S. Jeong, "Dna encoding for splice site prediction in large dna sequence," in *International Conference on Database Systems for Advanced Applications*. Springer, 2013, pp. 46–58.

[39] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M. S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W. C. Clark *et al.*, "The dynamic n 1-methyladenosine methylome in eukaryotic messenger rna," *Nature*, vol. 530, no. 7591, p. 441, 2016.

[40] A. Khan, S. Shah, F. Wahid, F. G. Khan, and S. Jabeen, "Identification of microrna precursors using reduced and hybrid features," *Molecular BioSystems*, vol. 13, no. 8, pp. 1640–1645, 2017.

[41] M. Kabir and M. Hayat, "irspot-gaensc: identifing recombination spots via ensemble

classifier and extending the concept of chou's pseaac to formulate dna samples," *Molecular genetics and genomics*, vol. 291, no. 1, pp. 285–296, 2016.

[42] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "iss-psednc: identifying splicing sites using pseudo dinucleotide composition," *BioMed research international*, vol. 2014, 2014.

[43] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro *et al.*, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific reports*, vol. 5, p. 10312, 2015.

[44] M. J. Iqbal, I. Faye, B. B. Samir, and A. Md Said, "Efficient feature selection and classification of protein sequence data in bioinformatics," *The Scientific World Journal*, vol. 2014, 2014.

[45] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, 2015, pp. 1200–1205.

[46] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "Svm-rfe: selection and visualization of the most relevant features through non-linear kernels," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–18, 2018.

[47] L. de Paula, A. S. Soares, T. W. de Lima, and C. J. Coelho, "Feature selection using genetic algorithm: an analysis of the bias-property for one-point crossover," in *Proceedings of the 2016 on genetic and evolutionary computation conference companion*. ACM, 2016, pp. 1461–1462.

[48] B. Wutzl, K. Leibnitz, F. Rattay, M. Kronbichler, M. Murata, and S. M. Golaszewski, "Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness," *PloS one*, vol. 14, no. 7, 2019.

[49] G. I. Webb and Z. Zheng, "Multistrategy ensemble learning: Reducing error by

combining ensemble learning techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 8, pp. 980–991, 2004.

[50] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[52] J. Ramírez, J. Górriz, R. Chaves, M. López, D. Salas-Gonzalez, I. Alvarez, and F. Segovia, "Spect image classification using random forests," *Electronics letters*, vol. 45, no. 12, pp. 604–605, 2009.

[53] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[54] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[55] R. Desrosiers, K. Friderici, and F. Rottman, "Identification of methylated nucleosides in messenger rna from novikoff hepatoma cells," *Proceedings of the National Academy of Sciences*, vol. 71, no. 10, pp. 3971–3975, 1974.

[56] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, "N 6-methyladenosine marks primary micrornas for processing," *Nature*, vol. 519, no. 7544, pp. 482–485, 2015.

[57] S. Xiang, Z. Yan, K. Liu, Y. Zhang, and Z. Sun, "Athmethpre: A web server for the prediction and query of mrna m 6 a sites in arabidopsis thaliana," *Molecular BioSystems*, vol. 12, no. 11, pp. 3333–3337, 2016.

[58] I. Nazari, M. Tahir, H. Tayara, and K. T. Chong, "in6-methyl (5-step): Identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc," *Chemometrics and Intelligent Laboratory Systems*, vol. 193, p. 103811, 2019.

[59] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mcpred: a sequence-based meta-predictor for accurate dna 4mc site prediction using effective

feature representation," *Molecular Therapy-Nucleic Acids*, vol. 16, pp. 733–744, 2019.

[60] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[62] H. Alashwal, S. Deris, and R. M. Othman, "Comparison of domain and hydrophobicity features for the prediction of protein-protein interactions using support vector machines," *International Journal of Information Technology*, vol. 3, no. 1, pp. 1305–2403, 2006.

[63] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with chou's pseudo amino acid composition," *Journal of theoretical biology*, vol. 384, pp. 78–83, 2015.

[64] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear b-cell epitopes using amino acid pair antigenicity scale," *Amino acids*, vol. 33, no. 3, pp. 423–428, 2007.

[65] W. Chen, H. Tang, and H. Lin, "Methyrna: a web server for identification of n6-methyladenosine sites," *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 3, pp. 683–687, 2017.

[66] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "irna-methyl: Identifying n6-methyladenosine sites using pseudo nucleotide composition," *Analytical biochemistry*, vol. 490, pp. 26–33, 2015.

[67] I. Nazari, M. Tahir, H. Tayara, and K. T. Chong, "in6-methyl (5-step): Identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc," *Chemometrics and Intelligent Laboratory Systems*, vol. 193, p. 103811, 2019.

[68] Y. Zhuang, H. Liu, X. Song, Y. Ju, and H. Peng, "A linear regression predictor for identifying n6-methyladenosine sites using frequent gapped k-mer pattern," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 673–680, 2019.

[69] K. Liu and W. Chen, "imrm: a platform for simultaneously identifying multiple kinds of rna modifications," *Bioinformatics*, 2020.

[70] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mcpred: a sequence-based meta-predictor for accurate dna 4mc site prediction using effective feature representation," *Molecular Therapy-Nucleic Acids*, vol. 16, pp. 733–744, 2019.

[71] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[72] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear b-cell epitopes using amino acid pair antigenicity scale," *Amino acids*, vol. 33, no. 3, pp. 423–428, 2007.