Bio-Medical Materials and Engineering 24 (2014) 3779–3785
DOI 10.3233/BME-141207
IOS Press

Characterization and prediction of mRNA alternative polyadenylation sites in rice genes

Xiaohui Wu[a], Chuang Zhao[a] and Yaru Su[b,*]
[a]*Department of Automation, Xiamen University, Xiame 361005, China*
[b]*Forensic Science Division, Department of Fujian Provincial Public Security, Fuzhou 361003, China*

Abstract. Polyadenylation [poly(A)] of mRNA is a critical step during gene expression, which plays an important role in the termination of transcription. Prediction of poly(A) sites can help identify 3' ends of genes and improve genome annotation. Due to the limited knowledge of poly(A) signals in plants, predictive modeling of poly(A) sites in agricultural crops remains challenging. Recent studies have uncovered widespread occurrences of alternative poly(A) (APA) sites in intron and coding sequence (CDS), whereas the study on the prediction of these APA sites is scarce. In this study, four feature representation methods, involving a position weight matrix, the k-gram frequency, core hexamers, and a transition matrix, were adopted to characterize poly(A) signals surrounding APA sites. The classification model was built to predict each group of APA sites. Experimental results showed that this model was effective in the identification of APA sites located in different genomic regions, with a compromise between sensitivity and specificity higher than 87%. Compared with previous model PASS rice, accuracies for the prediction of APA sites in 3'-UTR, intron and CDS were enhanced by 5%, 7%, and 27%, respectively. This model will contribute to genetic engineering by enabling researchers to control poly(A) site selection.

Keywords: Rice, polyadenylation, alternative polyadenylation, prediction, classification model

1. Introduction

mRNA polyadenylation is a critical cellular process during gene expression which adds poly(A) tails to mature mRNAs [1]. Polyadenylation has been shown to be associated with several human diseases such as breast cancer [2]. The poly(A) site marks the end of the mature mRNA, which can be used to identify genes and define gene boundaries. Therefore, prediction of poly(A) sites gives insights into gene structures and improve genome annotation. To date, a number of approaches have been proposed for predicting poly(A) sites in yeast, human, etc. An early approach by Graber and coworkers used a hidden Markov model (HMM) to predict poly(A) sites in yeast [3]. Cheng et al. used a support vector machine (SVM) to predict human poly(A) sites [4]. Akhtar et al. developed POLYAR, which classified poly(A) sites into three groups containing different forms of poly(A) signal and predicted poly(A) sites in each group [5]. In addition, several attempts have also been made for the prediction of

*Corresponding author: Yaru Su, Forensic Science Division, Department of Fujian Provincial Public Security, Fuzhou 361003, China. Tel.: +86 18606020083; Fax: 0592-2580258; E-mail: yarusu@gmail.com.

poly(A) sites in plants. Ji et al. proposed the generalized HMM to prediction poly(A) sites in Arabidopsis [6,7]. Classification models using Bayesian network and combined classifiers were also employed for poly(A) site prediction in Arabidopsis and *Chlamydomonas reinhardtii*, respectively [8–10]. The limited knowledge of poly(A) signals in plants makes it much more challenging for predicting plant poly(A) sites. Moreover, recent studies have uncovered widespread occurrences of alternative poly(A) (APA) sites in intron and coding sequence (CDS) [11,12]. Unfortunately, these types of APA sites were considered as control data for many identification methods [7,9,10], which would definitely affect the prediction accuracy. Till now, the study on the prediction of unconventional APA sites in intron and CDS is scarce.

As a model plant, rice is one of the most important crops in the world. Genomic study of rice facilitates genomic researches of other grain crops. In this study, several feature representation methods, involving a position weight matrix (PWM), the k-gram frequency, core hexamers, and a transition matrix, were employed to characterize represented poly(A) signals surrounding poly(A) sites in 3'-UTR, intron, and CDS. Then, the cost-sensitive meta subset evaluator and information gain method [13] were adopted to select a relative best feature space. Finally, the classification model integrating several classifiers was built to predict constitute poly(A) sites in 3'-UTR and alternative poly(A) sites in intron and CDS.

2. Materials and methods

2.1. Datasets

The rice poly(A) site dataset was from the previous study [12]. Poly(A) sites were divided into three groups (3'-UTR, intron, and CDS) based on their locations. 500, 200, and 100 sequences were randomly selected from group 3'-UTR, intron, and CDS as positive training datasets, respectively. The same numbers of sequences were randomly selected from the rest of data to construct the test positive dataset. These sequences are all of length 180 nt, with poly(A) site at the 150^{th} position [12,14]. The control dataset was from three types of sequences without poly(A) sites, including 5'-UTR, intergenic, and randomly generated sequences using the Markov chain (MC) [7]. The control training dataset for group 3'-UTR includes 100 5'-UTR, 200 intergenic, and 200 MC sequences. The control training dataset for group intron includes 40 5'-UTR, 80 intergenic, and 80 MC sequences. The control training dataset for group CDS includes 20 5'-UTR, 40 intergenic, and 40 MC sequences. The control test dataset has the same number of sequences as the control training dataset.

2.2. Representation of sequence features in poly(A) signal regions

To identify poly(A) sites, it is required to characterize sequence features surrounding poly(A) sites. Several poly(A) signal regions have been reported in previous study, including FUE (Far Upstream Element, -150~-35 nt from the poly(A) site), NUE (Near Upstream Element, -35~-10 nt from the poly(A) site), and CE (Cleavage Element, -15~10 nt around the poly(A) site) [12,14]. The widely used motif recognition tool MEME [15] was used to search motifs with length 4 to 8 nt. Identified motifs were then visualized as sequence logos (Figure 1). Next, various feature representation methods, including a PWM [15], the k-gram frequency, core hexamers, and a transition matrix, were adopted to convert these sequence features into numeric vectors (Figure 1).

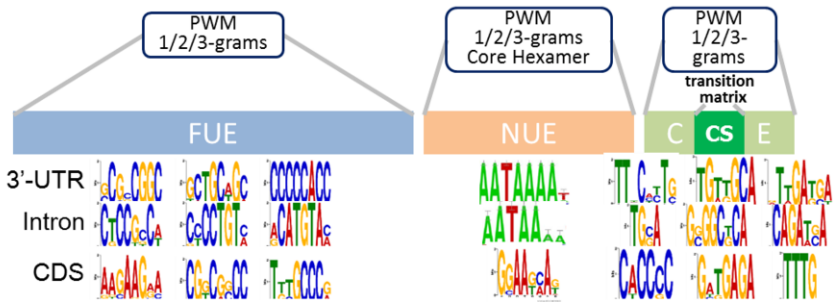

Fig. 1. Motifs discovered and feature representation methods of different groups of poly(A) sites. CS: cleavage site.

2.3. Position weight matrix

Each identified motif (Figure 1) can be denoted by a PWM. In this matrix, rows correspond to the four nucleotide bases (A, T, C, and G) and columns correspond to the positions of the motif. The value of each item in the matrix is the observed frequency (probability) of each possible nucleotide. Given a study region, a PWM value can be calculated for each motif in this region. Given a region with length L nt and a motif with length k nt, the PWM value at the i^{th} position is:

$$S_i = \frac{1}{k}\sum_{i}^{i+k-1} P_{ib} \tag{1}$$

where b indexes the bases and P_{ib} is the PWM value of the base at the i^{th} position.

2.4. K-gram frequency

K-gram is a short oligomer with length k nt. 1, 2 and 3-grams (mononucleotide, dinucleotide, and trinucleotide) were considered as candidate patterns. For regions of FUE, NUE, and CE, frequencies of 4 mononucleotides, 16 dinucleotides, and 64 trinucleotides were calculated, respectively. The initial feature space contains 252 k-grams (k=1, 2, and 3). Attribute selection methods were used to further select effective feature subset. The cost-sensitive meta subset evaluator [13] was first employed to select a candidate subset. Next, each candidate attribute in this subset was measured by information gain [13]. Candidate attributes with information gain higher than 0.1 were remained. Finally, 5, 7, and 7 k-grams were obtained for groups of 3'-UTR, intron, and CDS, respectively.

2.5. Core hexamers

In plants, NUE is much conserved than other regions, where AATAAA is the most dominant signal in this region [14]. Therefore, core hexamers with a high number of occurrences in NUE region were selected for more comprehensive characterization of this signal region. Given an NUE region, k hexamers appearing in this region were obtained first. Next, corresponding frequencies were calculated. Then these frequencies were normalized as values between 0 and 1 Eq. (2). The higher the value is, the higher the probability of the occurrence of this hexamer in NUE region is.

$$f_{ik} = [f_{ik} - \min(f_k)]/[\max(f_k) - \min(f_k)] \tag{2}$$

Here, f_{ik} is the value of the i^{th} hexamer, and f_k denotes values of k hexamers. Finally, hexamers with values higher than 0.25 were selected as core hexamers.

2.6. Transition matrix

In plants, the composition of nucleotide base of cleavage site (CS) is YA dinucleotide (Y=C or T) [14]. To represent this dinucleotide structure in a more effective way, the CS dinucleotide (299-300 nt in a 400 nt sequence, or 149-150 in a 180 nt sequence) was represented using transition matrix. To this end, the first order Markov chain was employed to calculate the probability transition of CS dinucleotide. The probability of the first base of CS dinucleotide was calculated as $P1 = \{P_A, P_T, P_C, P_G\}$, and that of the dinucleotide was calculated as $P2 = \{P_{AA}, P_{AT}, \dots, P_{GC}, P_{GG}\}$. For a given sequence, the transition probability of CS dinucleotide XY is $P1 \times P2$.

2.7. Model training and testing

A classification based model was built to predict poly(A) sites. For each group, the corresponding positive and control training datasets were used for training. A training model integrating eight classifiers was built to generate a model file for each group of poly(A) sites. The eight classifiers are Random Forests, Bayes Network learning, Naive Bayes classifier, sequential minimal optimization algorithm, AdaBoost M1 method, alternating decision tree, normalized Gaussian radial basis function network, and logistic regression model [10,13]. For a training dataset, eight training models $M_1, \dots, M_8$ were generated using the eight classifiers. Then the respective test dataset was tested using all the training models. After testing, two values are generated for each input sequence in the test dataset for each training model, denoting the probabilities of the 150^{th} position in this sequence being or not being a poly(A) site (true / false). The final true and false probabilities of an input sequence are:

$$P_T = \sum_{i=1}^{i=8} T_i \,;\; P_F = \sum_{i=1}^{i=8} F_i \tag{3}$$

where T_i is the true probability of the i^{th} training model, F_i is the false probability. Then the final true probability is normalized between 0 and 1.

3. Results

Two widely used assessment criteria were employed to evaluate the prediction performance, sensitivity (Sn) and specificity (Sp). Positive sequences with poly(A) sites were used to calculate Sn. Negative sequences without any poly(A) site were used to calculate Sp. Similar to previous study [9,10], a tolerance was allowed in the evaluation of Sn and Sp. For a given tolerance M nt, if at least one position within $\pm M$ of the true poly(A) site is predicted as true, then it is a true positive prediction. If no position within $\pm M$ of the false poly(A) site is predicted as true, then it is a true negative prediction.

Poly(A) sites in 3'-UTR, intron, and CDS were tested using the corresponding model, respectively. As shown in Figure 2A, all the three groups show overall high Sn (Sn>90%), indicating the selected features and the training models are effective in the prediction of positive sequences. Particularly, for CDS poly(A) sites which are lack of dominant AATAAA signal, the performance is still very high, suggesting the importance of combining features from multiple signal regions. When no margin is al

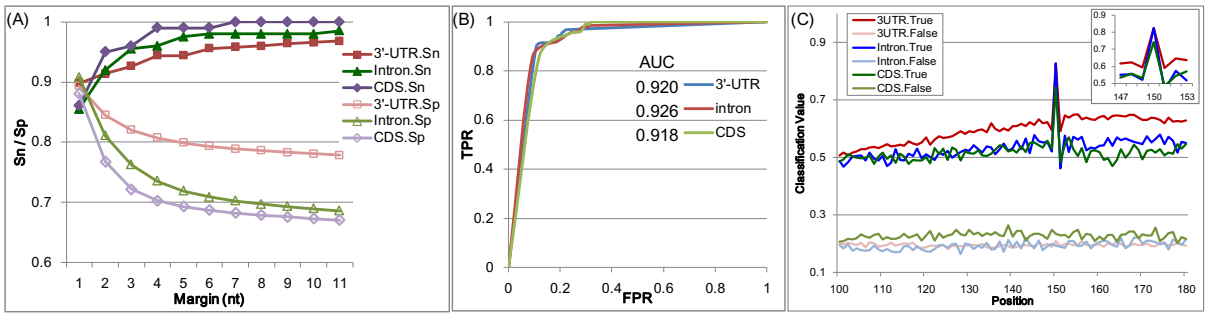

Fig. 2. Prediction results for each group of poly(A) sites. (A) Sn and Sp as a function of the margin; (B) ROC curves and AUC values; (C) average prediction probabilities of the test dataset.

lowed (M = 0), the Sn of 3'-UTR group is the highest among the three groups, reflecting the location of poly(A) sites in 3'-UTR is the most precise. Although the 3'-UTR group shows relatively lower Sn profile (M > 0), Sp of 3'-UTR group is apparently higher than that of intron and CDS, demonstrating the effectiveness of features and models for distinguishing true 3'-UTR poly(A) site and the false one. With the increase of the margin, Sn values are increased and Sp values are decreased. The choice of a margin for assessing the predictions is a trade-off between Sn and Sp. For a margin of 5 or higher, Sn values approach 95%, however, Sp values drop below 80%. Therefore, at lower margin, the prediction becomes increasingly robust, as the possibility of random occurrences drops.

ROC (Receiver Operating Characteristic) is a fundamental tool for diagnostic test evaluation. The area under the ROC curve (AUC) is a metric of how well a model can distinguish between two groups. Based on the Sn and Sp values of each group, the corresponding ROC curve was plotted (Figure 2B). Although Sn and Sp profiles among the three groups are distinct, their ROC curves show no difference with AUC values all above 0.90. In this study, sequences of length 400 nt with poly(A) site at the 300th position were also tested. For each positive and negative dataset of 400 nt sequences, probabilities of each positions of all sequences in each dataset were averaged. As shown in Figure 2C, probabilities of control datasets are all very small (less than 0.3) and the probability curves are very flat without any spike. In contrast, probabilities of positive datasets are much higher and a local spike was observed at the position of poly(A) site, indicating the classification model is capable of distinguishing between true and false poly(A) sites. The profiles of positive CDS and intron datasets are similar, which may due to that both poly(A) sites from CDS and intron are unconventional. Taking together, this result demonstrates that the sequence features defined in this study can highlight the poly(A) site region and significantly enhance the accuracy of the detection of poly(A) sites.

PASS and PASS_rice based on GHMM were developed to predict poly(A) sites in Arabidopsis and rice in previous studies [6,7,12]. In this study, PASS_rice was adopted to predict the test datasets for comparison. It is noteworthy that PASS_rice aims at the prediction of poly(A) sites in 3'-UTR and may not be fully suitable for the prediction of poly(A) sites in intron and CDS. However, all the three groups of poly(A) sites were tested using PASS_rice for a more comprehensive comparison. As shown in Figure 3A, although Sn curves are similar among the three groups, the Sp of CDS is apparently lower than that of intron and 3'-UTR, reflecting that PASS_rice is not suitable to be directly applied on CDS poly(A) sites. AUC values from PASS_rice are much lower than those from our model (Figure 3B vs. Figure 2B). Especially for CDS group, the AUC value from PASS_rice is only 0.657, which is 36% lower than that from our model. Such a sharp difference of AUC values for CDS group between two models may due to that our model enables the specific feature selection, model training

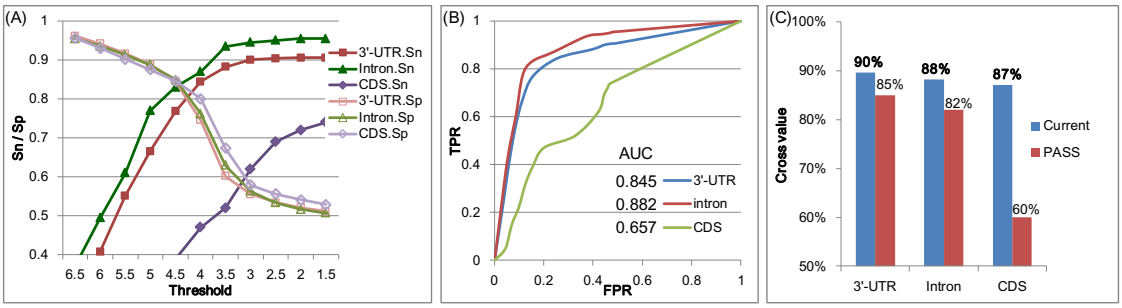

Fig. 3. Results from PASS_rice and the comparison between two models. (A) Sn and Sp from PASS_rice method; (B) ROC curves and AUC values from PASS_rice; (C) comparison of cross values of Sn and Sp between PASS_rice and our model.

and testing for CDS poly(A) sites, whereas parameters of PASS_rice were trained from 3'-UTR poly(A) sites. Cross values of Sn and Sp from these two models were also compared (Figure 3C). Compared to PASS_rice, prediction performances of our model for 3'-UTR, intron, and CDS poly(A) sites are enhanced by 5%, 6%, and 27%, respectively. Again, these results demonstrate the effectiveness of our model in the prediction of alternative poly(A) sites from different genomic regions.

4. Conclusion

Poly(A) signals in plants are much less conserved than those in mammals, leading to the challenge in the prediction of plant poly(A) sites. As data accumulate, there is still no computational method for the prediction of APA sites located in intron and CDS. The classification based model developed in this study is the first attempt to predict alternative poly(A) sites in rice. Various feature representation methods could be employed for the characterization of different poly(A) signals. Different classifiers could be adopted for model training and testing. The prediction results demonstrated the efficacy of the proposed model. The average prediction performance was enhanced by 5% to 27% compared with the previous GHMM-based prediction tool PASS rice (Figure 3C). The flexibility of this model was also demonstrated by the high prediction performance of APA sites in intron and CDS.

Because poly(A) sites define the ends of mature mRNA, the proposed model will be useful in genome annotation. This model will yield reliable poly(A) site candidates, providing important clues for relevant biological studies. In particular, it can be adopted to predict potential poly(A) sites for lowly expressed genes that are not normally found using EST experiments, or genes without known poly(A) sites. Additionally, this model can also be used to identify unconventional APA sites rather than constitute poly(A) sites in 3'-UTR. APA contributes to the transcriptome diversity, generating isoforms with different 3'-ends or coding capacity. Our study of APA sites in rice will provide insights into regulatory mechanisms of mRNA polyadenylation and promote the potential use of APA manipulation to reduce crop disease. This model will also be useful in genetic engineering by enabling researchers to control poly(A) site selection in designing transgenes.

Acknowledgement

This work was funded by the National Natural Science Foundation of China (No. 61201358), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research

Fund for the Doctoral Program of Higher Education of China (No. 20120121120038), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: No. 2013121025).

References

[1] A.A. Mueller, T.H. Cheung and T.A. Rando, All's well that ends well: alternative polyadenylation and its implications for stem cell biology, Current Opinion in Cell Biology **25** (2013), 222–232.
[2] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen and A. Xu, Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing, Genome Research **21** (2011), 741–747.
[3] J.H. Graber, C.R. Cantor, S.C. Mohr and T.F. Smith, Genomic detection of new yeast pre-mRNA 3'-end-processing signals, Nucleic Acids Research **27** (1999), 888–894.
[4] Y. Cheng, R.M. Miura and B. Tian, Prediction of mRNA polyadenylation sites by support vector machine, Bioinformatics **22** (2006), 2320–2325.
[5] M.N. Akhtar, S.A. Bukhari, Z. Fazal, R. Qamar and I. Shahmuradov, POLYAR, a new computer program for prediction of poly(A) sites in human sequences, BMC Genomics **11** (2010), 1–10.
[6] G. Ji, X. Wu, J. Zheng, Y. Shen and Q.Q. Li, Modeling plant mRNA poly(A) sites: Software design and implementation, Journal of Computational and Theoretical Nanoscience **4** (2007), 1365–1368.
[7] G. Ji, J. Zheng, Y. Shen, X. Wu, R. Jiang, Y. Lin, J.C. Loke, K.M. Davis, G.J. Reese and Q.Q. Li, Predictive modeling of plant messenger RNA polyadenylation sites, BMC Bioinformatics **8** (2007), 1–15.
[8] G. Ji, X. Wu, J. Huang and Q.Q. Li, Implementation of a classification-based prediction model for plant mRNA poly(A) sites, Journal of Computational and Theoretical Nanoscience **7** (2010), 927–932.
[9] G. Ji, X. Wu, Y. Shen, J. Huang and Q.Q. Li, A classification-based prediction model of messenger RNA polyadenylation sites, Journal of Theoretical Biology **265** (2010), 287–296.
[10] X. Wu, G. Ji and Y. Zeng, In silico prediction of mRNA poly(A) sites in Chlamydomonas reinhardtii, Molecular Genetics and Genomics **287** (2012), 895–907.
[11] X. Wu, M. Liu, B. Downie, C. Liang, G. Ji, Q.Q. Li and A.G. Hunt, Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation, Proceedings of the National Academy of Sciences, USA **108** (2011), 12533–12538.
[12] Y. Shen, G. Ji, B. J. Haas, X. Wu, J. Zheng, G.J. Reese and Q.Q. Li, Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation, Nucleic Acids Res. **36** (2008), 3150–3161.
[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD explorations newsletter **11** (2009), 10–18.
[14] J.C. Loke, E.A. Stahlberg, D.G. Strenski, B.J. Haas, P.C. Wood and Q.Q. Li, Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures, Plant Physiol. **138** (2005), 1457–1468.
[15] T.L. Bailey, N. Williams, C. Misleh and W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Research **34** (2006), W369–W373.