Identification of Plant Messenger RNA Polyadenylation Sites Using Length-Variable Second Order Markov Model

Guoli Ji, Huanghui Zhang, Xiaohui Wu
Department of Automation
Xiamen University, Xiamen, China 361005
glji@xmu.edu.cn, zhhu245@163.com,
xhuister@xmu.edu.cn

Meishuang Tang
Modern Educational Technical and Practical Training
Center, Xiamen University
Xiamen, China 361005
tangms@xmu.edu.cn

***Abstract*—In this paper we adopted a length-variable second order Markov model to identify plant messenger RNA poly(A) sites, and provided a common method that only relies on the experimental sequences. The efficacy of our model is showed up to 92% sensitivity and 79% specificity. This method is particularly suitable for the prediction of the poly(A) site which is lack of biological priori knowledge and has poor conservative signal characteristic, as well as for the identification of the alternative poly(A) sites in different genetic regions. Compared with other algorithms, generalized hidden Markov model needed the signal distributions and AdaBoost required the construction of signal features around the sites, our model is more versatile.**

Keywords-polyadenylation site; length-variable second order Markov model; biological priori knowledge.

I. INTRODUCTION

In eukaryotic cells, precursor messenger RNA (mRNA) needs to transcribe and undergo processing events, including 5' capping, intron splicing, and 3' end polyadenylation (poly(A)), then becomes mature and functional mRNA. Therein, the 3' end polyadenylation processing affects mRNA stability, translatability, and nuclear-to-cytoplasmic export, and plays an important role in other cellular and disease mechanisms [1-4].

Nowadays many researchers have worked on the feature and the prediction of poly(A) sites of eukaryotic mRNA sequences in different species, such as in human [5-8], yeast [9,10], Caenorhabditis [11] and Arabidopsis [12-14]. According to the studies of these organisms, it is found that consensus motifs or features surrounding the poly(A) sites in animals are more conservative than those in plants. Previous researches show that the poly(A) signals in plant mRNA sequences are more dispersed, diverse and complex [15]. There is little conservation in cis-elements and many 3' ends of plant mRNAs do not have the most conservative poly(A) signal AAUAAA, which can only be found in about 10% of Arabidopsis genes [16,17], but in up to 58% humans genes [7]. Moreover, multiple poly(A) sites often exist in plant [17], for example, 14 different 3' processing sites were found in a tobacco chloroplast mRNA binding protein coding gene [8]. Therefore, it is a more challenging task to identify poly(A) sites in plants then in animals.

Current protocols to recognize plant poly(A) sites rely heavily on expressed sequence tags (ESTs) with a poly(A) tract. And, there are several poly(A) sites databases available [8,10]. However, many poly(A) sites cannot be accurately identified because differential expression, mis-annotation and incomplete EST data [5]. Moreover, current protocols to identify poly(A) sites mostly aims at the animals and humans [6-11]. Though there are some methods to predict plant poly(A) site, such as GHMM (Generalized Hidden Markov Model) [12] and AdaBoost [13], they still have limitations. Both the models based on GHMM and AdaBoost largely depend on the signal features surrounding the poly(A) sites, known as FUE, NUE and CS [15], and pre-process of the mRNA sequences such as feature generation, feature selection and feature standardization. Thus, it is very difficult to identify the poly(A) sites without such pre-processes using these two models. In this paper, we adopted length-variable second order Markov model (LVMM2) [18] to identify mRNA poly(A) sites in model plant *Arabidopsis thaliana* through probabilistic and statistical computing of upstream and downstream nucleotide sequences of the poly(A) sites to obtain the corresponding transition probability matrixes. Compared with the GHMM and AdaBoost models, LVMM2 only depends on the experimental sequences to build the recognition model, but does not require any priori knowledge. The results show that our model not only achieves higher identification accuracy, but also is less time-consuming.

II. METHODS

In this paper, we adopted LVMM2 to predict plant poly(A) sites. The LVMM2 was first proposed and used in prediction of splice sites in human genome, which containing two different models. LVMM2 can be constructed with little pre-processing of the experimental sequences to predict the candidate sites by the likelihood at each position [18]. We borrowed the idea of this model to predict poly(A) sites in that it is especially suitable for processing sequences with little priori knowledge. However, the identification of plant poly(A) sites is huge different from identification of the splice sites, because the

This work was supported by the National Natural Science Foundation of China (No. 60774033), Specialized Research Fund for the Doctoral Program of Higher Education of China (No.20090121110022), the Fundamental Research Funds for the Central Universities of Xiamen University (No. 2011121047), and Xiamen University's National 211 Project 3rd period (No. 0630-E62000).

species is different and there are insufficient consensus motifs or features surrounding the poly(A) sites. Here, we took the advantage of this algorithm for the prediction of poly(A) sites in Arabidopsis. More details about the model method are in the rest of this section.

A. Markov model

Here we consider the gene sequence with a poly(A) site as a Markov process. The sequence is denotes as $L_1 L_2 \cdots L_H$, the state set is $M = \{A, G, T, C\}$. Then, for a given training set, the Markov transition probability is constructed by the following formula:

$$R_i(L_i) = \frac{X(L_{i-k} \cdots L_i)}{X(L_{i-k} \cdots L_{i-1})} \tag{1}$$

Where $X(L_{i-k} \cdots L_i)$ means the number occurrences of $L_{i-k} \cdots L_i$ in the training set.

Assuming that we get a K^{th} order Markov model MM, the likelihood of sequence $l_1, l_2 \cdots, l_H$ is

$$r(l_1 l_2 \cdots l_H \mid MM) = \prod_{i=1}^{H} R_i(L_i) \tag{2}$$

B. Description of length-variable second order Markov model

Given a sequence, it is predicted according to its likelihood under the Markov model which is considered as a classifier. The ratio of the likelihood under true site model and the false site model is compared, then a threshold *hd* is applied to predict poly(A) sites. According to the relevant experiments, however, it was found that there are little consensus motifs or features surrounding plant poly(A) sites [17]. Thus, it is insufficient to precisely identify poly(A) sites only relying on a certain length, and the interception of a long sequence may bring about redundant or irrelevant attributes which leads to reduce the recognition performance of the model [18]. For these reasons, we adopted the length-variable second order Markov model to predict Arabidopsis poly(A) sites.

In LVMM2, the di-nucleotides occurrence probability in training sequences surrounding the poly(A) sites is statistically analyzed by first, then the ratio of the likelihood (denoted as I) under the true site model and false site model is compared to predict the poly(A) sites. In the test model, two different thresholds are set, T_thr and F_thr ($T\_thr > 1.0$, $F\_thr \le 1.0$), then a candidate sequence is predicted. The sequence is predicted as containing poly(A) site if $I > T\_thr$, while not containing poly(A) site if $I < F\_thr$. If $F\_thr \le I \le T\_thr$, it indicates that the sequence is not significant to be classified, so more features are needed, then we extend the length and repeat the previous process to recalculate the I until $I > T\_thr$ or $I < F\_thr$. If $F\_thr \le I \le T\_thr$ is still even reaching the maximum length of the sequence, then the candidate sequence is classified by $I > (T\_thr + F\_thr)/2$ standing for true and $I \le (T\_thr + F\_thr)/2$ as false.

C. Prediction algorithm

Given the length of training sequence is $L = L_U + 2 + L_D$ (the length of poly(A) site is 2), we construct the models from the true and false data set. Because mRNA sequences is instable, the sequences used in our experiment are DNA sequences, so we used T instead of U in RNA, but this does not affect the analysis. For upstream model, the state set is denoted as $M^U = \{A, T, G, C\}$ and the random process is $\{L_t \mid t = 1, 2, \cdots, H_U\}$ ($L_t \in \{A, T, G, C\}$), then according to the second order Markov model, we construct the model R^{TU} (true upstream model). Similar to the models R^{TD} (true downstream model), R^{FU} (false upstream model) and R^{FD} (false downstream model). Supposing a test sequence $L = L^U_{h_U}, L^U_{h_U-1} \ldots L^U_1, [YA], L^D_1, \ldots L^D_{h_D-1}, L^D_{h_D}$ with default upstream length h_U ($h_U \le H_U$) and downstream length h_D ($h_D \le H_D$), its likelihood *I(L)* is calculated as follows:

$$R^T(L) = \prod_{i=1}^{h_U} R_i^{TU}(L_i) \times \prod_{j=1}^{h_D} R_j^{TD}(L_j) \tag{3}$$

$$R^F(L) = \prod_{i=1}^{h_U} R_i^{FU}(L_i) \times \prod_{j=1}^{h_D} R_j^{FD}(L_j) \tag{4}$$

$$\begin{aligned} I(L) = \frac{R^T(L)}{R^F(L)} &= \frac{\prod_{i=1}^{h_U} R_i^{TU}(L_i) \times \prod_{j=1}^{h_D} R_j^{TD}(L_j)}{\prod_{i=1}^{h_U} R_i^{FU}(L_i) \times \prod_{j=1}^{h_D} R_j^{FD}(L_j)} \\ &= \prod_{i=1}^{h_U} \frac{R_i^{TU}(L_i)}{R_i^{FU}(L_i)} \times \prod_{j=1}^{h_D} \frac{R_j^{TD}(L_j)}{R_j^{FD}(L_j)} \end{aligned} \tag{5}$$

The sequence model is shown in Figure.1, the "italic nucleotides" are the default length of the models and the "bold di-nucleotides" means the poly(A) site.

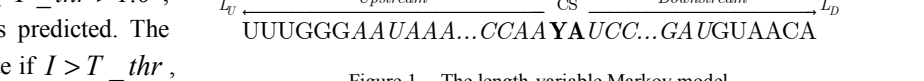

Figure 1. The length-variable Markov model

If the value of *I(L)* is neither greater than T_thr nor less than F_thr, we need to extend the upstream or downstream length by the following formulas:

$$\begin{aligned} &DIRECT = 0, \\ &(UpInc > 1.0 \,\&\, DownInc > 1.0 \,\&\, UpInc \geq DownInc) \\ &\| \,(UpInc < 1.0 \,\&\, DownInc < 1.0 \,\&\, UpInc \leq DownInc) \end{aligned} \quad (6)$$

$$\begin{aligned} &DIRECT = 1, \\ &(UpInc > 1.0 \,\&\, DownInc > 1.0 \,\&\, UpInc < DownInc) \| \| \\ &(UpInc < 1.0 \,\&\, DownInc < 1.0 \,\&\, UpInc > DownInc) \end{aligned} \quad (7)$$

$$\begin{aligned} &DIRECT = NonLeft, \\ &(UpInc \geq 1.0 \,\&\, DownInc \leq 1.0) \| \| \\ &(UpInc \leq 1.0 \,\&\, DownInc \geq 1.0) \end{aligned} \quad (8)$$

Here:

$$UpInc = R^{TU}_{h'_U+1}(L_{h'_U+1}) / R^{FU}_{h'_U+1}(L_{h'_U+1}) \quad (9)$$

$$DownInc = R^{TD}_{h'_D+1}(L_{h'_D+1}) / R^{FD}_{h'_D+1}(L_{h'_D+1}) \quad (10)$$

Where the '0' indicates extending upstream and '1' shows extending downstream, '&' stands for 'AND', '||' means 'OR'. *NonLeft* means selecting the different direction of last time. When $h'_U = H_U$, the downstream need extend, and if $h'_D = H_D$ means update the upstream length [18].

D. Model default length and thresholds

The initial length is important in our model in that it directly affects the identification accuracy and experiment time. However, if we have no idea about the characteristics surrounding the poly(A) sites or the construction of the dominant region, we can set the default length parameters through different combinations of upstream and downstream length to meet different requirements in the modeling process. This is also the reason why LVMM2 model is less dependent on the priori knowledge.

In addition to initial length, our models also contain two other parameters: thresholds *T_thr* and *F_thr*. In this paper, we tested several groups of *T_thr* and *F_thr* according to the experimental standards to determine the optimal value for *T_thr* and *F_thr*.

III. RESULTS

A. Datasets

The dataset with real poly(A) sites was extracted from GenBank (release 85.0, Dec. 08, 2003), with a total of 8160 ESTs (8K dataset) [12]. These sequences were trimmed into 400 nt in length, with known poly(A) sites at the 301st position, including 301 nt upstream and 99 nt downstream of the authentic poly(A) site. And several control datasets without any poly(A) site were truncated from the regions of Introns，CDS (coding sequences) and 5'UTR (5' untranslated region), respectively. All these sequences are 400nt in length.

To test the efficiency of our model, we generated one training dataset to estimate the optimal model parameters, and another test dataset to test our model. The training dataset includes the true dataset containing 4000 sequences randomly selected from the 8K dataset, and the false dataset with 20,000 sequences randomly selected from the control datasets. For the test dataset, we randomly selected 4000 sequences from the rest of 8K dataset to form true dataset and another 20,000 sequences from the control datasets to form the false dataset.

B. Performance standards

To evaluate the performance of our experiments, we adopt two common standards: sensitivity (Sn) and specificity (Sp):

$$Sn = \frac{TP}{TP + FN} \quad (11)$$

$$Sp = \frac{TN}{TN + FP} = 1 - \frac{FP}{TN + FP} \quad (12)$$

TP (true positive) means the number of actual sites predicted correctly. FN (false negative) is the number of true sites not predicted correctly. TN (true negative) is the number of false sites predicted correctly. FP (false positive) represents the number of negative sites predicted as true sites. The value of Sn represents the fraction of the true poly(A) sites which can be identified correctly, while Sp means the proportion of false sites correctly predicted.

C. Results

To show the effectiveness of our model, first we tested our model on the true dataset and false dataset mentioned above and used ROC curves to evaluate the results. ROC curve provides a comprehensive evaluation for the accuracy of the prediction, using the false positive rate (FPR, 1-Sp) as the horizontal axis and the true positive rate (TPR, Sn) as the vertical axis. In this paper, we set (1-Sp) of a fixed *F_thr* 0.2 with *T_thr* from 1.5 to 6.0 as the horizontal axis and Sn as the vertical axis. As shown in Figure. 2, different length combinations were used, such as '6+7', '16+17', '46+32', '100+50' ('6+7' means the default upstream length + default downstream length).

It was found that the identification performance would be better with the increasing of default upstream and downstream length. Therefore, based on the previous analysis, we tested our model with thresholds *T_thr*=1.5 and *F_thr*=0.2. The results are shown in Table I.

From Table I, it showed that the value of Sn increased at the beginning, but became fluctuated when the default upstream length reached 150nt. This indicated that it might bring about redundant or irrelevant attributes to the model when the upstream length is more than 150nt. In addition, the maximum upstream length of experimental sequence is 300 nt. Therefore, the longer of the initial length is, the shorter length to be extended when the candidate sequence cannot be significant identified, which would affect the recognition performances. For these considerations, we selected the upstream 150 nt and downstream 30 nt to further test our model. Here, we tested the prediction accuracy with several *F_thr* (0.2, 0.3, 0.4, 0.5, 0.6), and obtained the ROC curves of each *F_thr* by updating the value of *T_thr* (Figure.3).

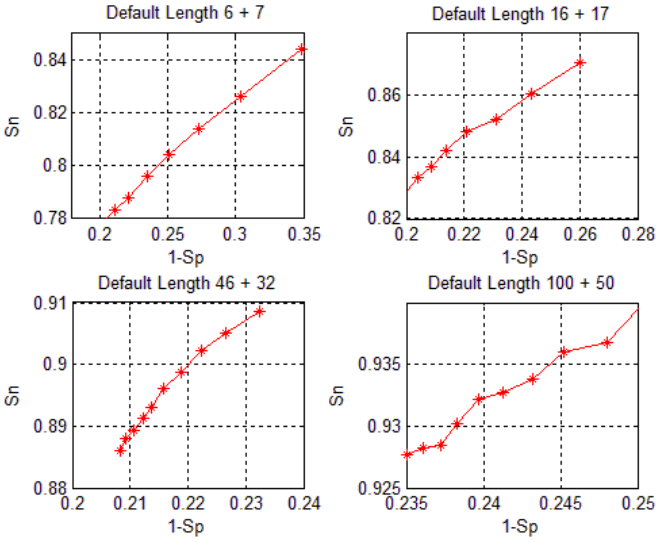

Figure 2. ROC curves with different length

TABLE I. RESULTS OF DIFFERENT LENGTH

Left + Right	Sn	1-Sp
50+20	0.9240	0.2169
50+30	0.9264	0.2186
50+40	0.9278	0.2208
100+20	0.9430	0.2272
100+30	0.9424	0.2298
100+40	0.9430	0.2326
150+20	0.9438	0.2433
150+30	0.9444	0.2444
150+40	0.9462	0.2464
200+20	0.9404	0.2431
200+30	0.9426	0.2444
200+40	0.9448	02455

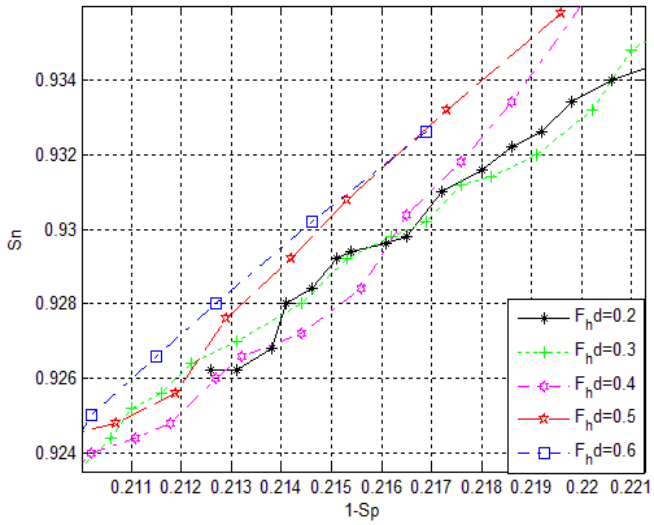

Figure 3. The ROC curves of different *F_thr*

Through analysis of these results and the relevant information mentioned above, finally we set the default upstream length as 150 nt and downstream as 30 nt, the thresholds *T_thr*=3.0 and *F_thr*=0.5, respectively, as our model parameters. Then, we applied our model on the test dataset using these parameters. Our model achieved high identification performance with Sn=92.2% and Sp=79.3%. Here, we compared our results with one previous AdaBoost based poly(A) recognition model [13]. Using AdaBoost [13], the Sn was 70% and Sp was about 89% in different control datasets. In contrast, our LVMM2 model achieved much higher Sn by 22%, which means much more of the true poly(A) sites were identified correctly by LVMM2 than by AdaBoost. Though, our Sp is about 10% lower than that of AdaBoost, this may due to that we didn't use different control datasets but the combined one to test our model. In fact, it is not practical to use different control datasets to calculate Sp, since normally the user has little priori knowledge about his sequence. Therefore, our model is more universal, because it did not rely on the type of the sequence, or the genetic region the sequence may be located. Using the combination of randomly selected sequences from different genetic regions as the false dataset made our model be especially suitable for the prediction of long gene sequence, with satisfied Sn at expense of reducing Sp a bit.

IV. CONCLUSIONS

In this paper, our model not only achieved high identification performance, but also is more flexible than previous poly(A) site recognition models, like AdaBoost [13] and GHMM [12]. AdaBoost model [13] is required to extract the features of the datasets by different methods, like K-gram nucleotide sequence pattern, Z-Curve, PSSM-based CIS score, etc. For new datasets, it needed to recalculate the weight of every sample set. While the GHMM [12] is needed the signal distributions according to priori knowledge to make each genetic region framework and constituted a fixed pattern. Once the features are changed or new signal characteristics are added, we have to reconstruct the model step by step. In

contrast, our Length-variable second order Markov Model can achieve higher identification accuracy without any priori knowledge or pre-processing the experiment data. Thus, our model is especially suitable for the prediction of plant mRNA poly(A) sites which are lack of biological priori knowledge and conservative signal characteristic, as well as for the identification of the alternative poly(A) sites in different genetic regions.

With the development of biotechnology, given more information accumulated from the biological experiments in the future, it is easy to modify the LVMM2 model to improve the prediction accuracy, such as changing the order of model, weighting the different genetic regions, or combine the other algorithms to form a dynamic model framework to optimize our identification model.

ACKNOWLEDGMENT

We are grateful to Quanwei Zhang for his helpful suggestions.

REFERENCES

[1] G. Edwalds-Gilbert, K.L .Veraldi and C. Milcarek, "Alternative poly(A) site selection in complex transcription units: mean to an end?," Nucleic Acids Res, vol. 25, pp. 2537-2561, 1997.

[2] J. Zhao, L. Hyman and C. Moore, "Formation of mRNA 3' ends in eukaryotes: mechanism and regulation," Microbiol. Mol. Biol. Rev, vol. 63, pp. 405-445, 1999.

[3] B. Conne, A. Stutz and J.D. Vassalli, "The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology," Nat. Med, vol. 6, pp. 637-641, 2000.

[4] N.H. Gehring, U. Frede, G. Neu-Yilik, et al, "Increased efficiency of mRNA 3'end formation: a new genetic mechanism contributing to hereditary thrombophilia," Nature Genet, vol. 28, pp. 389-392, 2001.

[5] E. Beaudoing, S. Freier, J.R. Wyatt, J.M Claverie, D, Gautheret. (2000) "Patterns of variant polyadenylation signal usage in human genes," Genome Research, vol. 10, pp. 1001-1010, 2000.

[6] H. Liu, H. Han and J. Li, (2003) "An In-Silico Method for Prediction of Polyadenylation Signals in Human Sequences," Genome Informatics, vol. 14, pp. 84-93, 2003.

[7] J. Hu, S.L. Carol, W. Jeffrey, B. Tian, "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation," RNA Society, vol. 11, pp. 1485-1493, 2005.

[8] Y. Cheng, R.M. Miura and B. Tian, "Prediction of mRNA polyadenylation sites by support vector machine," Bioinformatics, vol. 22, no. 19, pp. 2320-2325, 2006.

[9] J.H. Graber, C.R. Cantor, S.C. Mohr, T.F. Smith, "Genomic detection of new yeast pre-mRNA 3' -end-processing signals," Nucleic Acid Research, vol. 27, no. 3, pp. 888-894, 1999.

[10] J.H. Graber, G.D. McAllister and T.F. Smith, "Probabilistic prediction of Saccharomyces cerevisiae mRNA 3' -processing sites," Nucleic Acid Research, vol. 30, no. 8, pp. 1851-1858, 2002.

[11] A. Hajarnavis, I. Korf and Durbin R, "A probabilistic model of 3' end formation in Caenorhabditis elegans," Nucleic Acids Research, vol. 32, no. 11, pp. 3392-3399, 2004.

[12] G. Ji, J. Zheng, Y. Shen, et al., "Predictive modeling of plant messenger RNA polyadenylation sites," BMC Bioinformatics, vol. 8, no. 43, 2007.

[13] G. Ji, D. Zou, J. Zheng, Q.Q. Li, "An AdaBoost Algorithm for the identification of Arabidopsis Messenger RNA Polyadenylation Sites," ICISE, 1: 3579-3582, 2009.

[14] G. Ji, J. Zheng, D. Zou, et al., "Recognition of Plant mRNA Polyadenylation Sites Based on High-Dimensional Space Points' Covering Method," ISISE, 2: 616-620, 2008.

[15] H.M. Rothnie, "Plant mRNA 3'-end formation,". Plant Molecular Biology, vol. 32, pp. 43-61, 1996.

[16] C.P. Joshi, "Putative polyadenylation signals in nuclear genes of higher plants: a compilation and analysis," Nucleic Acid Research, vol. 15, pp. 9627-9640, 1987.

[17] J.C. Loke, E.A. Stahlberg, D.G. Strenski, et al., "Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures," Plant Physiology, vol. 138, pp. 1457-1468, 2005.

[18] Q. Zhang, Q. Peng, Q. Zhang, Y. Yan, K.K. Li, J. Li, "Splice sites prediction of human genome using length-variable Markov model and feature selection," Expert Systems with Applications, vol. 37, pp. 2771-2782, 2010.