A Novel Genome-wide Polyadenylation Sites Recognition System Based on Condition Random Field

Jiuqiang Han, Shanxin Zhang, Jun Liu and Ruiling Liu*

***Abstract*—Polyadenylation including the cleavage of pre-mRNA and addition of a stretch of adenosines to the 3'-end is an essential step of pre-mRNA processing in eukayotes. The known regulatory role of polyadenylation in mRNA localization, stability, and translation and the emerging link between poly(A) and disease states underline the necessary to fully characterize polyadenylation sites. Several artificial intelligence methods have been proposed for poly(A) sites recognition. However, these methods are suitable to small subsets of genome sequences. It is necessary to propose a method for genome-wide recognition of poly(A) sites. Recent efforts have found a lot of poly(A) related factors on DNA level. Here, we proposed a novel genome-wide poly(A) recognition method based on the Condition Random Field (CRF) by integrating multiple features. Compared with the polya_svm (the most accurate program for prediction of poly(A) sites till date), our method had a higher performance with the area under ROC curve(0.8621 versus 0.6796). The result suggests that our method is an effective method in genome wide poly(A) sites recognition.**

I. Introduction

During the mature process of mRNA and most lncRNA, polyadenylation or poly(A) generated by RNA polymerase II is a fundamental event and plays a crucial functional role in RNA stability, nuclear export efficiency, and subsequent translation [1]. Alternation of the process can perturb cell growth and is associated with multiple human diseases including cancer [2]. Recent studies have revealed that alternative polyadenylation (APA) is pervasive with mRNAs and lncRNAs [3, 4]. Recent reports also provide unexpected and attractive evidence that genes often convert their expression toward the shorter 3'UTR isoforms that correspond to truncated versions of the canonical long isoforms in proliferating or cancer cells [2, 5]. Thus, identifying poly(A) sites or PAS in genome is one of the essential problems in understanding the mechanisms of the regulation process.

Recently, the advance of experimental technology largely improves the accuracy of detecting the locations of poly(A) sites. New approaches such as Direct RNA sequencing or DRS, 3P-Seq, PAS-Seq, 3'READ and others are beginning to be used to near-completely identify poly(A) sites[2-4, 6, 7]. These methods can find out the poly(A) sites in a high throughput and with a high accuracy, but they have several limitations: 1) these methods could detect the poly(A) sites of only one tissue in one experiment and it is expensive, for different tissues and different species, plenty of experiment are required; 2) the internal priming problem could cause false positives recognition and add false poly(A) sites. Thus, computational methods are needed for poly(A) sites identification as complement.

Plenty of efforts have been made to predict poly(A) sites by searching patterns around poly(A)sites in genome. Position weight matrix (PWM) is the most widely used model to represent and recognize poly(A) sites[8, 9]. However, PWM alone is not discriminative enough and will predict lots of false positives, due to the fact that the motifs are very short and often degenerated. Recently, various approaches have been proposed to reduce false positives by integrating information of the sequences composition, such as the *k-mer* and chemical features [10, 11]. However, these methods only suit to a subset of the whole genome. Recently, a lot of works have found plenty of polyadenylation related factors, such as transcription factors[12], histones markers[13], miRNA target sites[14, 15], and RNA binding proteins[16, 17] *etc*.

In this paper, we present a novel method based on Conditional Random Field (CRF) to identify poly(A) sites. CRF was introduced to bioinformatics area recently, such as gene prediction[18, 19], transcription factors recognition[20], and presented promising results. CRF can integrate information from different sources and capture complex dependency. Therefore it is an ideal framework for poly(A) sites prediction. Different types of features, the Position Weight Matrix (PWM), miRNA Target sites position, CpG Island position, transcription factors, histone markers and RNA binding proteins have been integrated in the method. The system diagram of our method was shown in Figure 1.

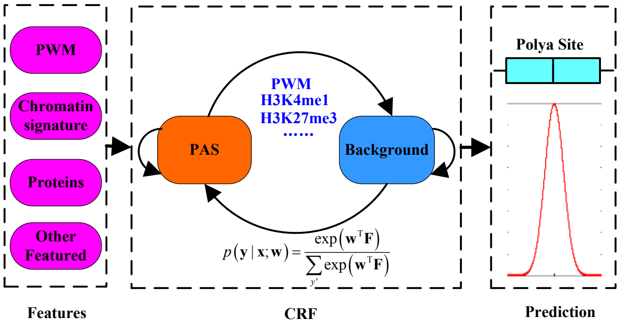
PWM
Chromatin signature
Proteins
Other Featured
PWM
H3K4me1
H3K27me3
……
PAS
Background
$p(\mathbf{y}\mid\mathbf{x};\mathbf{w})=\frac{\exp(\mathbf{w}^{T}\mathbf{F})}{\sum_{y'}\exp(\mathbf{w}^{T}\mathbf{F})}$
Polya Site
Features
CRF
Prediction

Figure 1. The flow chart of our CRF-based Poly(A) sites prediction method

This work was supported by grants from National Natural Science Foundation of China (No. 61105021), Ph.D. Program Foundation of the Ministry of Education of China (No. 20110201110010), China Postdoctoral Science Foundation (No. 2011M501442) and the Fundamental Research Funds of Xi'an Jiaotong University (xjj2013020).

J. Han (e-mail: jqhan@xjtu.edu.cn), S. Zhang (e-mail: shanxin.zh@gmail.com), J. Liu (e-mail: 525023952@qq.com) and R. Liu (e-mail: meggie@xjtu.edu.cn) are with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China.

*Corresponding author, phone:+86-029-82668665-175, E-mail: meggie@mail.xjtu.edu.cn

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

II. Materials and Methods

A. Datasets

The annotated Poly(A) sites data were directly downloaded from the GENCODE Project sites[21]. The histone information, CpG Island sites, miRNA target sites and ChIP-Seq data or RIP-Seq and the sequences of human genome (Refseq hg19) were downloaded from UCSC Genome Browser [22].

B. Generating gold-standard PAS dataset and feature matrix

Four kinds of features,including the Position Weight Matrix (PWM) of 15 *cis*-element used in polya_svm [9], the distance to miRNA and to CpG Island, histone markers (8 distinct histone modifications), and transcript factors (Pabpc1, Elav1, and Pol2) are used.

To generate gold-standard dataset on the binned genome, the "Peak-centric" method was used [20]. First, we divided genome into 200*nt* bins. Then, we assigned bins that were overlapped with the centers of PASs as positive ones and other bins as negative ones. Similar method was employed to acquire a feature matrix. The PWM scores of a certain bin was the maximal PWM score in their corresponding range inside the bin. Then, for other features, the value corresponding to a histone or transcript factor of a certain bin was set to 1 if that bin overlapped with one peak and 0 otherwise. In the case of miRNA site and CpG Island proximity, if bins overlapped with 500nt upstream and downstream that region, their values were set to 1; otherwise, they were 0.

C. CRF-based Poly(A) Site Annotation Tool

Our CRF-based method has been proposed to predict polyadenylation sites by aggregating information from different sources. In this method, a genome was first divided into 200*nt* bins. Then, the conditional probability-like score of a given observation sequence was computed as follows

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\lambda}) = \frac{\exp\left(\sum_{t=1}^{L}\sum_{k=1}^{K}\lambda_k f_k(\gamma_t, \gamma_{t-1}, t, \mathbf{x})\right)}{\sum_{y'}\exp\left(\sum_{t=1}^{L}\sum_{k=1}^{K}\lambda_k f_k(\gamma'_t, \gamma'_{t-1}, t, \mathbf{x})\right)}, \qquad (1)$$

where $\mathbf{y}$ is the label sequence, $\mathbf{x}$ is the observed genomic sequence. f_k is the k^{th} feature function and λ_k is the corresponding weight. The function f_x can be an arbitrary function on $\mathbf{x}$. And here y' is any label sequence. For label sequence, the possible values are 0 and 1.

D. Feature Design

Similar to the transcription factor binding site prediction tool CTF [20], We designed four kinds of feature functions to catch patterns involved in features as well. The first type is PWM scoring function. The second type is indicator function, which examines the occurrence of a feature. The third type which is used to capture co-occurring features aims at the co-occurrence of two features. In addition, a kind of feature functions that captures patterns in adjacent bins is used as complements.

In our method, for different types of feature functions, different function templates were designed. Let $\mathbf{x}$ be a feature matrix, then $x_{i,j}$ is the value of i^{th} feature in the j^{th} bin in the genome. $\mathbf{y}$ represents the label sequence and y_j is the label of the j^{th} bin. $I\{conditions\}$ is an indicator function. Its value is 1 if and only if all conditions are true. The first kind of feature functions is used for PWM, and it is defined as

$$f(\gamma_j, \gamma_{j-1}, j, \mathbf{x}) = x_{1,j} I\{\gamma_j = u\}, \qquad (2)$$

where u indicating the label of that bin is 0 or 1. The second kind of functions which is used to check the occurrence of features is defined as

$$f(\gamma_j, \gamma_{j-1}, j, \mathbf{x}) = I\{\gamma_{j-1} = u, \gamma_j = v, x_{i,j} = 1\}, \qquad (3)$$

Where both u and v represent labels. The third type aims to the co-occurrence of two features and its definition is

$$f(\gamma_j, \gamma_{j-1}, j, \mathbf{x}) = I\{\gamma_{j-1} = u, \gamma_j = v, x_{i,j} = 1, x_{i',j'} = 1\}, \qquad (4)$$

where i and i' corresponds to two features. At last, feature functions that capture patterns in adjacent bins as a complement for above feature functions are used, and these are defined as

$$f(\gamma_j, \gamma_{j-1}, j, \mathbf{x}) = I\{\gamma_{j-1} = u, \gamma_j = v, x_{i',j-1} = 1, x_{i,j} = 1\}, \qquad (5)$$

and

$$f(\gamma_j, \gamma_{j-1}, j, \mathbf{x}) = I\{\gamma_{j-1} = u, \gamma_j = v, x_{i',j-1} = 1, x_{i'',j+1} = 1\}, \qquad (6)$$

where i and i' corresponds to two features, and u and v are labels.

E. Training

To estimate the parameter vector $\boldsymbol{\lambda}$, a Regularized Maximum Conditional Log Likelihood method are employed, which is defined as

$$\lambda_{ML} = \arg_{\lambda}\max\left(\ln\left(p(\gamma \mid x; \lambda)\right)\right), \qquad (7)$$

that is

$$\lambda_{ML} = \arg_{\lambda}\max\left(\sum_{t=1}^{L}\lambda_k f_k - \ln\left(Z(\mathbf{x})\right) - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2}\right), \qquad (8)$$

where $Z(\mathbf{x})$ is the partition function and $\|.\|$ is the L2 norm. The definition of $Z(\mathbf{x})$ is

$$Z(\mathbf{x}) = \sum_{y'}\exp\left(\sum_{t=1}^{L}\sum_{k=1}^{K}\lambda_k f_k(\gamma'_t, \gamma'_{t-1}, \mathbf{x})\right) \qquad (9)$$

The optimal weight vector λ was found by using liblbfgs (http://www.chokkan.org/software/liblbfgs/), an open source library for unconstrained minimization.

F. Prediction

To predict the label for a bin, the marginal probability of j^{th} bin was estimated as

$$s_j = p\left(\gamma_j = 1 \mid \mathbf{x}; \boldsymbol{\lambda}\right), \tag{10}$$

which is assigned as the score of each bin. Then, a threshold was set. The bins whose scores exceed the threshold will be assigned as polyadenylation sites, while the rest bins will be assigned as background.

The prediction method was implemented based on the framework of the CTF prediction tools.

G. Performance Evaluation

In order to evaluate the performance of our method, 2-fold cross-validation was used. we randomly divided the 22 autosomes and chromosomes X and Y into two groups. Then, one group was utilized as training set and the other as the test set. Area Under the Curve (AUC) of Receiver Operator Characteristic (ROC) curve was calculated to assess the performance.

We defined True Positives (TPs) as positive bins that were predicted as PASs and False Positives (FPs) as non-PASs bins that were predicted as PASs. Similarly, True Negatives (TNs) were negative bins predicted as non-TPASs. False Negatives (FNs) were defined as Negative bins predicted as positives. Then, True Positive Rate (TPR) was defined as the fraction of TPs predicted by a model in all positives. The fraction of FPs predicted by a model in all negatives was denoted by False Positive Rate (FPR).

We compared our method with polya_svm, a method using PWM scores of 15 *cis*-element as features and employing Support Vector Machine (SVM) as classifier, which is the most canonical and accurate method till date. In order to evaluate the two methods with the same criterion, the two methods are tested in the same bins.

III. Results and Discussion

DNA functional elements contained structural patterns and comparable information with the sequence when predicting poly(A) sites. As recent studies discovered that histone marker H3K36me3 could be a indicator for different kinds of polyadenylation sites [13]. In our work, 8 distinct histone markers were used: H3K27me3, H3K36me3, H3K79me2,H3K4me1, H3K4me2, H3K4me3, H3K9me3, and H4K20me1. Other features were the miRNA target sites and CpG Island proximity for the reason that most of the sequence surrounding the poly(A) sites are AT rich region and the miRNA in 3'UTR are most frequent before the polyadenylation site and the heptanucleotide TATTTAT to increase mRNA decay potency [15]. In addition, some transcript factors(Pabpc1, Elav1, and Pol2) are also used for poly(A) site prediction.

We got 15,478,399 bins with length of 200nt. The amount of bins containing PAS is 1,224,851 (about 7.91% of all the bins). The prediction results are shown in Figure 2. In the method, all features were included. because during the training process, unrelated features would get weights close to zero, thus we did not select features. Combining all features, the AUC value of our method is up to 0.8621.

To further evaluate our CRF-based method, we then compared our method with polya_svm– a SVM-based method. Polya_svm is an integrated method based on SVM and it predicts PASs based on scores of PWM. ROC curves of the two methods were shown in Figure2. Results showed that our method had better accuracy than polya_svm. The AUC of CRF was larger than AUC of SVM by 26.85% (0.8621 versus 0.6796). Next, we also compared the true positive rate (TPR) of all three methods at 10% false positive rate (FPR). Our method had the highest TPR (0.60), which was much better than TPR of SVM method (0.28). To sum up, our method outperformed existing methods in different metrics and it is an effective method in genome-wide polyadenylation sites prediction.

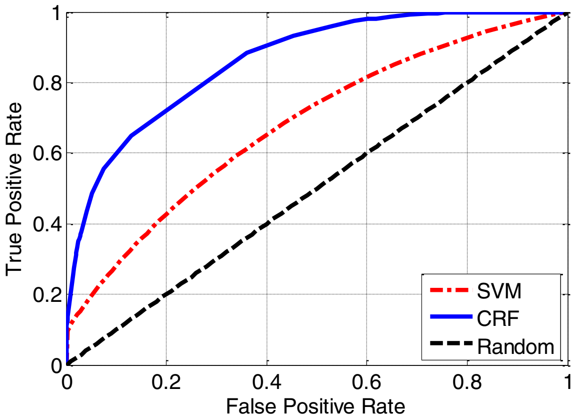

Figure 2. ROC curves of the polyadenylation sites prediction results based on Condition Random Field and Support Vector Machine methods. The X axis represents the False Positive Rate and the Y axis represents the True positive Rate. The Area Under Curve values of ROC for the CRF and SVM method are 0.8621 and 0.6796,respectively.

Although our method achieved a high accuracy, there are still much room for improvement. For example, in our current method, only the locations of the peaks of histone were considered. Continuous feature functions could be included in the future and that will contain the information of the shape and intensity. In addition, new features can be included in our method in a straightforward way due to the flexibility of CRF framework. We could integrate different poly(A) prediction methods results or decision values together as features in the future. For large genomes, it is necessary to accelerate the speed of our method and reduce the demand of RAM.

IV. Conclusions

In this paper, we present a novel integrative method to predict polyadenylation sites (PASs) by combining various features using conditional random field Our results showed that this CRF based method successfully integrated information from position weight matrix (PWM), miRNA target sites, CpG Island, distinct histone markers and transcript factors together. And it improved accuracy of PAS prediction greatly in total. When compared with existing representative tools, our method achieved obvious superior performance. The CRF based method is an effective novel integrative polyadenylation sites prediction system, and are with great potentials in discovering other functional elements.

ACKNOWLEDGMENT

The authors thank members of the laboratories of Dr. Dexing Zhong for helpful discussions. They also thank Mr. Zijun Xiong and Mr. Bo Li from BGI for advices on datasets collection.

REFERENCES

[1] C. Andreassi and A. Riccio, "To localize or not to localize: mRNA fate is in 3′ UTR ends," Trends in cell biology, vol. 19, pp. 465-474, 2009.

[2] Y. Lin, et al., "An in-depth map of polyadenylation sites in cancer," Nucleic Acids Research, vol. 40, pp. 8460-8471, Sep 2012.

[3] F. Ozsolak, et al., "Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation," Cell, vol. 143, pp. 1018-1029, Dec 10 2010.

[4] M. Hoque, et al., "Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing," Nature Methods, vol. 10, pp. 133-139, 2013-Feb 2013.

[5] C. Mayr and D. P. Bartel, "Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells," Cell, vol. 138, pp. 673-684, 2009.

[6] C. H. Jan, et al., "Formation, regulation and evolution of Caenorhabditis elegans 3 ' UTRs," Nature, vol. 469, pp. 97-114, Jan 6 2011.

[7] P. J. Shepard, et al., "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq," Rna-a Publication of the Rna Society, vol. 17, pp. 761-772, Apr 2011.

[8] M. N. Akhtar, et al., "POLYAR, a new computer program for prediction of poly(A) sites in human sequences," Bmc Genomics, vol. 11, Nov 19 2010.

[9] Y. Cheng, et al., "Prediction of mRNA polyadenylation sites by support vector machine," Bioinformatics, vol. 22, pp. 2320-2325, 2006.

[10] M. Kalkatawi, et al., "Dragon PolyA Spotter: predictor of poly (A) motifs within human genomic DNA sequences," Bioinformatics, vol. 28, pp. 127-129, 2012.

[11] F. Ahmed, et al., "Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies," In silico biology, vol. 9, pp. 135-148, 2009.

[12] R. Elkon, et al., "E2F mediates enhanced alternative polyadenylation in proliferation," Genome Biol, vol. 13, p. R59, 2012.

[13] C. Lee and L. Chen, "Alternative polyadenylation sites reveal distinct chromatin accessibility and histone modification in human cell lines," Bioinformatics, 2013.

[14] A. Wirsing, et al., "A systematic analysis of the 3'UTR of HNF4A mRNA reveals an interplay of regulatory elements including miRNA target sites," PloS one, vol. 6, pp. e27438-e27438, 2011 2011.

[15] A. Jacobsen, et al., "Signatures of RNA binding proteins globally coupled to effective microRNA target sites," Genome Research, vol. 20, pp. 1010-1019, 2010.

[16] E. de Klerk, et al., "Poly (A) binding protein nuclear 1 levels affect alternative polyadenylation," Nucleic Acids Research, vol. 40, pp. 9089-9101, 2012.

[17] M. Jenal, et al., "The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites," Cell, vol. 149, Apr 27 2012.

[18] D. DeCaprio, et al., "Conrad: gene prediction using conditional random fields," Genome Research, vol. 17, pp. 1389-1398, 2007.

[19] S. S. Gross, et al., "CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction," Genome Biology, vol. 8, p. R269, 2007.

[20] Y. He, et al., "CTF: a CRF-based transcription factor binding sites finding system," Bmc Genomics, vol. 13, p. S18, 2012.

[21] J. Harrow, et al., "GENCODE: The reference human genome annotation for The ENCODE Project," Genome Research, vol. 22, pp. 1760-1774, 2012.

[22] W. J. Kent, et al., "The human genome browser at UCSC," Genome Research, vol. 12, pp. 996-1006, 2002.