Bioinformatics, 31(10), 2015, 1671–1673
doi: 10.1093/bioinformatics/btv004
Advance Access Publication Date: 11 January 2015
Applications Note

Sequence analysis

PASPA: a web server for mRNA poly(A) site predictions in plants and algae

Guoli Ji[1,2], Lei Li[1], Qingshun Q. Li[3,4,5], Xiangdong Wu[1], Jingyi Fu[1], Gong Chen[1] and Xiaohui Wu[1,]*

[1]Department of Automation, [2] Innovation Center for Cell Biology and [3]Key Laboratory of the Ministry of Education on Costal Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, Fujian, China, [4]Department of Biology, Miami University, Oxford, OH, USA and [5]Rice Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian, China

*To whom correspondence should be addressed.

Associate Editor : John Hancock

Received on October 27, 2014; revised on December 10, 2014; accepted on December 30, 2014

Abstract

Motivation: Polyadenylation is an essential process during eukaryotic gene expression. Prediction of poly(A) sites helps to define the 3′ end of genes, which is important for gene annotation and elucidating gene regulation mechanisms. However, due to limited knowledge of poly(A) signals, it is still challenging to predict poly(A) sites in plants and algae. PASPA is a web server for **p**oly(**A**) **s**ite prediction in **p**lants and **a**lgae, which integrates many in-house tools as add-ons to facilitate poly(A) site prediction, visualization and mining. This server can predict poly(A) sites for ten species, including seven previously poly(A) signal non-characterized species, with sensitivity and specificity in a range between 0.80 and 0.95.

Availability and implementation: http://bmi.xmu.edu.cn/paspa

Contact: xhuister@xmu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Polyadenylation is one of the essential processes and a point of regulation for eukaryotic gene expression. This process involves the cleavage at the 3′end of pre-messenger RNA (pre-mRNA) and the addition of a poly(A) tail. A poly(A) site marks the end of a gene, the recognition of which is important to fully understand the functions of genes and to accurately annotate the genomes. Given the connection between polyadenylation and transcription termination, correct identification of poly(A) sites would also predict where transcription termination may occur. Moreover, alternative polyadenylation, the use of different poly(A) sites of a gene, has been recognized as an important gene expression regulator in eukaryotes.

A number of computational tools based on support vector machine or discriminant analysis have been developed to predict poly(A) sites or poly(A) signals in human, including Erpin (Gautheret and Lambert, 2001), POLYAR (Akhtar *et al.*, 2010), Dragon PolyA Spotter (Kalkatawi *et al.*, 2012), etc. Unlike human or animals where the dominant signal AAUAAA is found in >50% of poly(A) sites (Tian and Graber, 2012), no highly conserved poly(A) signal can be found in plants or algae (Xing and Li, 2011), there are subsequently fewer prediction models available. Our group developed a poly(A) site prediction tool called Poly(A) Site Sleuth (PASS) (Ji *et al.*, 2007) based on a generalized hidden Markov model (GHMM). PASS and its modified version have been applied on the poly(A) site prediction for Arabidopsis (Ji *et al.*, 2007) and rice (Shen *et al.*, 2008), respectively. Another classification based model called poly(A) site classifier (PAC) was developed for green alga *Chlamydomonas reinhardtii* (Wu *et al.*, 2012). Tzanis *et al.* (2011) also developed a classification-based command-line tool called polyA-iEP to predict poly(A) sites in Arabidopsis.

To our knowledge, most of the aforementioned tools were implemented as standalone programs with command-line interfaces, which limit their use among biologists. For example, PASS and PAC are Windows application tools and cannot run in other operation systems (OS). PolyA-iEP is a command-line tool specific to Arabidopsis. Here, we create the first open-access public web service for poly(A) site prediction in plants and algae (PASPA), integrating many in-house tools as add-ons to facilitate data visualization and mining. Our server covers 10 species, including rice, Arabidopsis, and *Medicago truncatula*, spikemoss *Selaginella moellendorffii*, moss P*hyscomitrella patens*, red alga *Cyanidioschyzon merolae*, two green algae *C.reinhardtii* and *Ostreococcus lucimarinus*, and two diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. Seven of these species are previously non-characterized in terms of their poly(A) signals and poly(A) sites.

2 Methods

2.1 Algorithm updates

PASS 2.0 is the core prediction program to predict poly(A) sites, which is based on the GHMM prediction model. An updated version of PASS with the following improvements is used. First, PASS is a Windows application tool implemented in Delphi language, limiting its use in command prompt or other OSes like Linux. Coded in C++, PASS 2.0 is executable in different OSes, as well as in command line. Second, the parameters for GHMM structure in PASS are fixed, e.g. the number of signal states, and therefore its parameters cannot be modified for other species whose poly(A) signals and number of signal states are different. In contrast, PASS 2.0 allows adjustment of the model parameters, such as length of each state, state type, and number of states, according to the characteristics of the surrounding sequences of target species to adapt to the respective poly(A) site profiles. Third, PASS 2.0 offers several probability calculation methods to calculate the state output probability of the GHMM structure, such as heterogeneous first-order Markov submodel and weights of signal patterns and their combination. Moreover, users can train their own GHMM parameter file for the prediction of poly(A) sites in a new species. Several Perl scripts are provided to facilitate better estimation of model parameters for any given species (described in Supplementary Section S1). Additionally, a different method is implemented to locate the exact positions of predicted poly(A) sites in PASS 2.0 rather than the approximate regions in PASS (described in Supplementary Section S2). We have trained our GHMM profiles for ten species and will continue to expand this list in the future when new poly(A) site data become available. However, if the species the user desired is not listed, similar species in the list can be chosen or our program can automatically choose the optimal model parameters according to the given sequence file.

2.2 PASPA server

PASPA is a web-based service consisting of a series of Perl scripts, a PHP web application, and PASS 2.0 for poly(A) site prediction, evaluation, and visualization. The input data for our server are a DNA sequence file in FASTA format, with the optional files for known poly(A) sites or model parameters. The output is the poly(A) site(s) predicted along the original target sequences, based on the poly(A) scores. The result pages contain three viewers (described in Supplementary Section S3). The first one is the list viewer in which the prediction results will be shown in a tabulated list. Users may also choose to visualize single nucleotide base compositions and poly(A) signals of the input sequences. The second one is the score viewer (Fig. 1a) in which all predicted and known poly(A) sites (if available) will be automatically anchored to the score series of the given sequence for visualization. With functions such as motif search, users can identify canonical poly(A) signals (e.g., AAUAAA or UGUAA) and their variants in the pattern viewer. The third viewer (Fig. 1b) is the EST-polyA viewer in which the input sequences are aligned to the genome to associate any authenticated polyadenylation evidence. There are four authentic poly(A) site datasets from whole-genome studies currently available, including Arabidopsis, rice, *M.truncatula*, and *C.reinhardtii* (Shen *et al.*, 2008; Wu *et al.*, 2011, 2014; Zhao *et al.*, 2014). Users can search known poly(A) sites in the vicinity of their aligned sequences and visualize both known sites and predicted ones along the genome. More details about the server implementation are described in Supplementary Section S3.

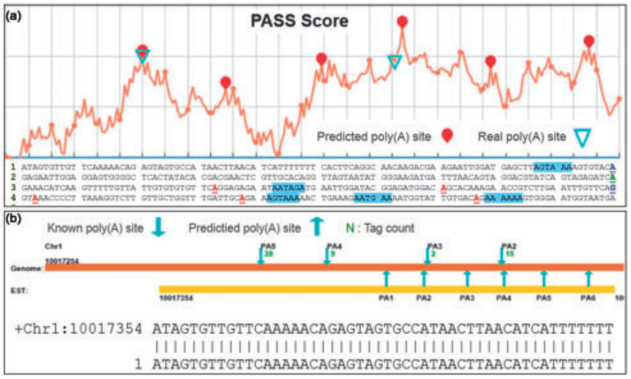
(a)
PASS Score
Predicted poly(A) site ●
Real poly(A) site ▽
1 ATAGTGTTGT TCAAAAACAG AGTAGTGCCA TAACTTAACA TCATTTTTTT CACTTCAGGC AACAAGACGA AGAATTGGAT GAGCTTAGTA AAAGTGTACA
2 GAGAATTGGA GGAGTGGGGC TCACTATACA CGACGAACTC GTTGCACAGG TTAGTAATAT GGGAAGATGA TTTAACAGTA GGACGTATCA GTAGAGATCA
3 GAAACATCAA GTTTTTGTTA TTGTGTGTGT TCAGGAGAGA ATAATAGATG AATTGGATAC GGAGATGGAC AGCACAAAGA ACGGTCTTGA ATTTGTTCAG
4 GTAAACCCCT TAAAGGTCTT GTTGCTGGTT TGATTGCAGA AAGTAAAAAC TGAAAGAATG AAAATGGTAT TGTGACAGAA AAAAGTGGGA ATGGTAATGA
(b)
Known poly(A) site
Predictied poly(A) site
N : Tag count
Chr1
10017254
Genome:
EST:
PA5
28
PA4
9
PA3
2
PA2
15
10017354
PA1
PA2
PA3
PA4
PA5
PA6
+Chr1:10017354 ATAGTGTTGTTCAAAAACAGAGTAGTGCCATAACTTAACATCATTTTTTT
1 ATAGTGTTGTTCAAAAACAGAGTAGTGCCATAACTTAACATCATTTTTTT

Fig. 1. Score viewer (**a**) and EST-polyA viewer (**b**) of PASPA

3 Results

For prediction purposes, the positive data set was generated from poly(A) sites in 3′ untranslated regions (3′-UTRs). Random sequences and sequences from 5′-UTRs or coding sequences were used as control datasets (described in Supplementary Section S4). Part of the sequences in the known poly(A) site datasets were randomly selected for model training. Part of the rest of the sequences from the positive ploy(A) datasets and the control datasets were used to estimate sensitivity (SN) and specificity (SP) (described in Supplementary Section S4). Similar to previous studies (Ji *et al.*, 2007; Shen *et al.*, 2008), a compromise between SN and SP was used to evaluate the model performance. Additional indexes including Matthew's correlation coefficient, the area under the receiver operating characteristic curve, F-measure were also employed for a more comprehensive evaluation of the performance. Cross value of SN and SP ranges from 0.80 to 0.95 (Table 1). The variation of performance among different species may due to the distinct set of poly(A) signals used in the respective species. In addition, compared with previous prediction tool, PASS for Arabidopsis and rice (Ji *et al.*, 2007; Shen *et al.*, 2008), PASPA obtains higher performance with better flexibility. More details about the model evaluation are given in Supplementary Section S5.

4 Conclusions

PASPA is the first web server for the prediction and visualization of poly(A) sites in plants and algae, with high prediction performance and good flexibility. It will be a valuable addition to the community for the study of polyadenylation, genome annotation, transcription, and gene expression.

Table 1. Cross values of SN and SP using the two control datasets

Plants	Ctrl1	Ctrl2	Algae	Ctrl1	Ctrl2
O.sativa	0.86	0.89	*C.merolae*	0.91	0.91
A.thaliana	0.90	0.88	*C.reinhardtii*	0.88	0.95
M.truncatula	0.83	0.87	*O.lucimarinus*	0.91	0.86
S.moellendorffii	0.89	0.89	*T.pseudonana*	0.93	0.87
P.patens	0.91	0.80	*P.tricornutum*	0.92	0.91

Ctrl1: cross value for control dataset of 5′-UTR or coding sequences.
Ctrl2: cross value for control dataset of random sequences.

Funding

The National Natural Science Foundation of China [61174161, 61201358 and 61375077], the Natural Science Foundation of Fujian Province of China [2012J01154], the specialized Research Fund for the Doctoral Program of Higher Education of China [20130121130004 and 20120121120038], the Key Research Project of Xiamen City of China [3502Z20123014], and the Fundamental Research Funds for the Central Universities in China Xiamen University [2013121025, 201412G009 and CBX2014007].

Conflict of Interest: none declared.

References

Akhtar,M.N. *et al.* (2010) POLYAR, a new computer program for prediction of poly(A) sites in human sequences, *BMC Genomics*, **11**, 646.

Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles, *J. Mol. Biol.*, **313**, 1003–1011.

Ji,G. *et al.* (2007) Predictive modeling of plant messenger RNA polyadenylation sites, *BMC Bioinformatics*, **8**, 43.

Kalkatawi,M. *et al.* (2012) Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences, *Bioinformatics*, **28**, 127–129.

Shen,Y. *et al.* (2008) Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation, *Nucleic Acids Res.*, **36**, 3150–3161.

Tian,B. and Graber,J.H. (2012) Signals for pre-mRNA cleavage and polyadenylation, *Wiley Interdiscip Rev. RNA*, **3**, 385–396.

Tzanis,G. *et al.* (2011) PolyA-iEP: A data mining method for the effective prediction of polyadenylation sites, *Expert Syst. Appl.*, **38**, 12398–12408.

Wu,X. *et al.* (2014) Genome-wide determination of poly(A) sites in *Medicago truncatula*: evolutionary conservation of alternative poly(A) site choice, *BMC Genomics*, **15**, 615.

Wu,X. *et al.* (2012) In silico prediction of mRNA poly(A) sites in *Chlamydomonas reinhardtii*, *Mol. Genet. Genomics*, **287**, 895–907.

Wu,X. *et al.* (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation, *Proc. Natl Acad. Sci. USA*, **108**, 12533–12538.

Xing,D. and Li,Q.Q. (2011) Alternative polyadenylation and gene expression regulation in plants, *Wiley Interdiscip Rev. RNA*, **2**, 445–458.

Zhao,Z. *et al.* (2014) Bioinformatics Analysis of Alternative Polyadenylation in Green Alga *Chlamydomonas reinhardtii* using transcriptome sequences from three different sequencing platforms. *G3: Genes/Genomes/Genetics*, **4**, 871–883.