BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 1 2012, pages 127–129
doi:10.1093/bioinformatics/btr602

Sequence analysis

Advance access publication November 15, 2011

Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences

Manal Kalkatawi[1,†], Farania Rangkuti[1,†], Michael Schramm[1,†], Boris R. Jankovic[1,†], Allan Kamau[1], Rajesh Chowdhary[2], John A. C. Archer[1] and Vladimir B. Bajic[1,*]

[1]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia and [2]Biomedical Informatics Research Center, MCRF, Marshfield Clinic, 1000 North Oak Avenue, Marshfield, WI 54449, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Recognition of poly(A) signals in mRNA is relatively straightforward due to the presence of easily recognizable polyadenylic acid tail. However, the task of identifying poly(A) motifs in the primary genomic DNA sequence that correspond to poly(A) signals in mRNA is a far more challenging problem. Recognition of poly(A) signals is important for better gene annotation and understanding of the gene regulation mechanisms. In this work, we present one such poly(A) motif prediction method based on properties of human genomic DNA sequence surrounding a poly(A) motif. These properties include thermodynamic, physico-chemical and statistical characteristics. For predictions, we developed Artificial Neural Network and Random Forest models. These models are trained to recognize 12 most common poly(A) motifs in human DNA. Our predictors are available as a free web-based tool accessible at http://cbrc.kaust.edu.sa/dps. Compared with other reported predictors, our models achieve higher sensitivity and specificity and furthermore provide a consistent level of accuracy for 12 poly(A) motif variants.

Contact: vladimir.bajic@kaust.edu.sa

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

Received on April 26, 2011; revised on September 30, 2011; accepted on October 26, 2011

1 INTRODUCTION

The polyadenylic acid tail or poly(A) tail is a stretch of A nucleotides added to RNA during the RNA processing mainly to protect the primary RNA stability (Bernstein and Ross, 1989). In mammals, the poly(A) tail is added close and downstream of the characteristic poly(A) signal, most often AAUAAA. The problem of prediction of poly(A) signals has received considerable attention. Since the distance of poly(A) signal from the poly(A) tail is approximately 10–30 nt (Beaudoing *et al.*, 2000), recognizing such tails in mRNA is relatively simple. A more challenging problem is to find a motif in primary genomic sequence that corresponds to poly(A) signal site in the transcribed RNA. The process of predicting poly(A) motifs in DNA depends on successfully identifying relevant properties of the surrounding sequence of such motifs. We now present a brief survey of reported work in this field so far. Statistical properties of nucleotide sequences were used, for example, to reveal putative poly(A) signals in yeast (Van Helden *et al.*, 2000) or in Arabidopsis (Ji *et al.*, 2010). A program PROBE was developed to identify *cis* elements that potentially play regulatory roles in mRNA polyadenylation (Hu *et al.*, 2005). Several tools are developed for predicting poly(A) motifs in human. Polyadq tool for predicting poly(A) motifs in a DNA sequence is reported in Tabaska and Zhang (1999) where sequences of 100 nt downstream of a candidate poly(A) motif were used to derive the feature set for prediction. The ERPIN program (Legendre and Gautheret, 2003) utilizes 300 nt flanking sequence upstream and downstream of the candidate poly(A) motifs. A method based on application of support vector machines (SVMs) was reported by Liu *et al.* (2003) in which 100 nt flanking sequence upstream and downstream around poly(A) candidate motifs were utilized. PolyApred system was introduced in Ahmed *et al.* (2009). The 100 nt flanking sequence upstream and downstream around candidate poly(A) motif sequence were utilized. A method POLYAR for recognition of polyadenylation sites is reported recently (Akhtar *et al.*, 2010). The reported results of these tools are summarized in Table 1, together with the performance of publicly available ones achieved on our datasets. In this study, we present a web-based tool that implements two types of predictive models, one based on Artificial Neural Networks (ANNs) and the other based on Random Forest (RF) (Breiman, 2001). Our models cover 12 main variants of human poly(A) motifs with accuracies from 82.06% to 94.4%.

2 METHODS

2.1 Datasets

We used human mRNA sequences and mapped 100 nt from their 3′ end back to the human genome applying stringent BLASTN matching criteria. Negative records were selected from human chromosome 21. Within candidate sequences, we selected those where the poly(A) motif is found at locations conforming to the distributions reported in Beaudoing *et al.* (2000). We flanked such poly(A) motifs by 100 nt upstream and 100 nt downstream, resulting in training sequences of 206 nt in length. Overall, 14 799 sequences for 12 motif variants can be found at http://cbrc.kaust.edu.sa/dps/code/DataToBuildModel.tar.gz. More details are given in Supplementary Material 1.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First authors.

Table 1. Accuracy of various poly(A) prediction tools

Tool	Results reported by authors	Results on our AATAAA dataset
Polyadq	MCC = 0.41–0.51	Se = 28.23% Sp = 83.88% Acc = 56.05%
Polya_SVM (Cheng *et al.*, 2006)	Se = 37.2–71.0% Sp = 74.6–96.7%	Se = 58.30% Sp = 64.42% Acc = 61.36%
Polyar	Se = 23.9–94.9% Sp = 14.7–66.4%	Se = 57.28% Sp = 49.69% Acc = 53.48%
Our Model (ANN)	Table 2	Se = 80.55% Sp = 83.57% Acc = 82.06%
Our Model (RF)	Table 2	Se = 86.10% Sp = 91.60% Acc = 88.90%
Polyah (Salamov, 1997)	MCC = 0.62	
ERPIN	Se = 56% Sp = 69–85%	
Polyapred	Se = 57.0% Sp = 75.8–95.7%	
Poly(A) Signal Miner (Liu *et al.*, 2003)	Se = 56.0–89.3% Sp = 67.5–93.3%	

2.2 Features and feature selection

Our model uses features from thermodynamic, compositional, statistical and other properties of nucleotides and polynucleotide sequences. The thermodynamic and structural properties of dinucleotides that we used were selected from Friedel *et al.* (2008). We also used electron–ion interaction potential (EIIP) of nucleotides (Veljkovic and Slavic, 1972). Finally, our models utilize scores from position weight matrices (PWMs) in the upstream and downstream regions of the poly(A) motifs. This process resulted in 274 features used (Supplementary Material 2).

2.3 The tool

For details of models see Supplementary Material 2. Our tool contains two types of predictors of poly(A) motifs, ANN-based and RF-based. The ANN models consist of an input, a hidden and an output layers. The output layer contains two neurons that predict if the input pattern corresponds to real or false poly(A) motif (the stronger wins). To mitigate overfitting, we deployed an early stopping method (Zang and Yu, 2005). The RF model is based on WEKA implementation (Hall *et al.*, 2009).

3 RESULTS

For performance we used sensitivity Se = TP/(TP + FN), specificity Sp = TN/(TN + FP) and accuracy Acc = (TP + TN)/(TP + TN + FP + FN), where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. We compared our results those of publicly available tools when applied to our datasets (Table 1). We tested on the only motif common to all tools (AATAAA). In Table 2, we report the performance of our ANN and RF-derived models on 12 poly(A) motifs. ANN model is trained on 50% of data and tested on the remaining 50% (training takes a long time so cross-validation is not applied). For the RF model, we achieved the best results using 100 trees without restricting maximal depth using nine random features per node. Model performance in 100-fold cross-validation is shown.

Table 2. Performance of ANN and RF methods for 12 poly(A) motifs

Varian	ANN mode		RF model	
	Se (%) Sp (%)	Acc (%)	Se (%) Sp (%)	Acc (%)
AAAAAG	94.57 85.44	90.01	93.2 95.6	94.4
AAGAAA	86.04 84.74	85.39	88.7 94.1	91.4
AATAAA	80.55 83.57	82.06	86.1 91.6	88.9
AATACA	91.71 88.39	90.05	87.3 92.5	89.9
AATAGA	95.18 93.37	94.27	86.7 91.3	89.0
AATATA	91.32 89.28	90.30	87.2 93.6	90.4
ACTAAA	89.85 89.49	89.67	85.0 91.1	88.1
AGTAAA	89.94 85.63	87.78	83.1 94.5	88.8
ATTAAA	83.71 83.96	83.84	85.2 92.6	88.9
CATAAA	91.56 91.98	91.77	83.5 92.4	88.0
GATAAA	88.75 91.66	90.20	87.9 92.5	90.2
TATAAA	92.30 86.20	89.25	86.1 94.2	90.1

4 CONCLUSION

We developed a web tool for the recognition of poly(A) motifs in human genomic DNA that demonstrates improved prediction accuracy over the existing publicly available poly(A) predictors. We hope that our tool will find good use in the studies of human gene properties.

Funding: This work is supported by the Base Research Funds of VBB at King Abdullah University of Science and Technology. The open access charges for this article are covered from the same fund.

Conflict of Interest: none declared.

REFERENCES

Ahmed,F. *et al.* (2009) Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. *In Silico Biol.*, **9**, 135–148.

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

Beaudoing,E. *et al.* (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.

Bernstein,P. and Ross,J.(1989) Poly(A), poly(A) binding protein and the regulation of mRNA stability. *Trends Biochem. Sci.*, **14**, 373–377.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Cheng,Y. *et al.* (2006) Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, **22**, 2320–2325.

Friedel,M. *et al.* (2008) DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.*, **37**, D37–D40.

Hall,M. *et al.* (2009) The WEKA Data Mining Software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

Hu,J. *et al.* (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, **11**, 1485–1493.

Ji,G. *et al.* (2010) A classification-based prediction models of mRNA polyadenylation sites. *J. Theor. Biol.*, **265**, 287–296.

Legendre,M. and Gautheret,D. (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics*, **4**, 7.

Liu,H. *et al.* (2003) An in-silico method for prediction of polyadenylation signals in human sequences. *Genome Inform.*, **14**, 84–93.

Salamov,A.A. and Solovyev,V.V. (1997) Recognition of 3'-processing sites of human mRNA precursors. *Comput. Appl. Biosci.*, **13**, 23–28.

Tabaska,J.E. and Zhang,M.Q. (1999) Detection of polyadenylation signals in human DNA sequences. *Gene*, **231**, 77–86.

Van Helden,J. *et al.* (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.

Veljkovic,V. and Slavic,I. (1972) Simple general model pseudopotential. *Phys. Rev. Lett.*, **29**, 105.

Zhang,T. and Yu,B. (2005) Boosting with early stopping: convergence and consistency. *Ann. Statist.*, **33**, 1538–1579.