A Simple Decision Rule for Recognition of Poly(A) Tail Signal Motifs in Human Genome

Hassan Abou Eisha, Igor Chikalov, and Mikhail Moshkov
Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
Email: {hassan.aboueisha, igor.chikalov, mikhail.moshkov}@kaust.edu.sa

Boris Jankovic
Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
Email: boris.jankovic@kaust.edu.sa

***Abstract*—Background is the numerous attempts were made to predict motifs in genomic sequences that correspond to poly (A) tail signals. Vast portion of this effort has been directed to a plethora of nonlinear classification methods. Even when such approaches yield good discriminant results, identifying dominant features of regulatory mechanisms nevertheless remains a challenge. In this work, we look at decision rules that may help identifying such features. Findings are we present a simple decision rule for classification of candidate poly (A) tail signal motifs in human genomic sequence obtained by evaluating features during the construction of gradient boosted trees. We found that values of a single feature based on the frequency of adenine in the genomic sequence surrounding candidate signal and the number of consecutive adenine molecules in a well-defined region immediately following the motif displays good discriminative potential in classification of poly (A) tail motifs for samples covered by the rule. Conclusions is the resulting simple rule can be used as an efficient filter in construction of more complex poly(A) tail motifs classification algorithms.**

***Index Terms*—poly (A) tails, decision rules, genomic sequences, machine learning, classification.**

I. Introduction

Polyadenilation is a process in which an mRNA molecule is terminated (appended) with a contiguous sequence of adenine molecules, primarily to improve the stability of the resulting molecule [1]. The starting position of this extension in mRNA is signaled by the sequence of nucleotides in mRNA, typically six nucleotides long, referred to as poly (A) signal. Large amount of work done so far relates to finding this signal in mRNA molecules, which in effect constrains the number of candidate signals. However, a closely related, but a more complex problem is to predict the location that corresponds to the polyadenilation signal in the primary genome that would be transcribed into the actual poly (A) signal in the resulting mRNA. This problem is important for functional analysis of genomic sequences. Generally, the sequences in the primary genomic sequence that correspond to actual poly(A) signal sequences (poly(A) tail motifs) are known, although their composition varies across species [2], [3]. In addition, there are typically several motifs indicating the true poly (A) signal, with varying degree of prominence. The analysis in this paper deals with polyadenilation in the human genome. The presence of these motifs in genomic sequence, however, is necessary, but not the sufficient condition for their translation to poly (A) signals. Therefore, this process can be reduced to the corresponding problem of binary classification of candidate motifs in the genomic sequence. Given a motif, together with its surroundings, the classification algorithm makes a prediction whether the given motif will correspond to a true polyadenilation signal or not. Many tools [4], [5], [6], [7], [8] were developed with a specific aim of performing such classification. In most cases these are implemented as neural networks, support vector machines, etc. utilizing model features (either statistical or physicochemical) derived from the sequences surrounding the motifs. It is, however, difficult to enunciate the dominant features implicated in the regulation of this process. For that reason, we build a predictive model based on decision trees to make identification of dominant features more feasible.

II. Methodology and Findings

The dataset used in this work was downloaded from [9] and contains 7370 positive and an equal number of negative samples in total for 12 variants of major human poly (A) signals. We therefore construct our samples to be used in building of the decision rules accordingly and each such sample represents the sequence of 100 nucleotides upstream of the 6-nucleotide motif (i.e. on the 5' side of the motif) and 100 downstream (on the 3' side

Manuscript received January 6, 2015; revised May 4, 2015.

doi: 10.12720/jait.6.2.71-74

of the motif). The sequences, including the motifs are thus 206 nucleotides in length. When analyzing these sequences, we use information in the nucleotide sequences surrounding the motifs, but not the nucleotide sequences within the motifs themselves. Part of the reason behind this strategy is that there are 12 polyadenilation signals in human genome which would firstly complicate analysis with uncertain payoff, and secondly it is questionable how much information from the motif is actually utilized by the polyadenilation regulatory mechanisms as the vast majority of poly(A) motifs in the primary genomic sequence are false (i.e. the corresponding sequences in mRNA are not poly(A) signals).

TABLE I. THRESHOLD VALUES AND CLASSIFICATION PERFORMANCE FOR 90% AND 95% CONFIDENCE FOR SEGMENTED DATASET

Adenine frequency range		Number of samples in bin	Target conf. level	Rule threshold		Rule coverage (samples per class per bin)			Rule coverage (of total)
min	max					Total	Pos	Neg	
0	0.295	Bin 1	90%	Pos	0.00400	2104	1894	210	77.42%
		Total 3818		Neg	0.02000	852	77	775	
		Positive 2488	95%	Pos	0.00048	1220	1159	61	47.72%
		Negative 1330		Neg	0.04700	602	30	572	
0.295	0.340	Bin 2	90%	Pos	0.00500	1770	1598	172	76.79%
		Total 3679		Neg	0.02800	1055	98	957	
		Positive 2151	95%	Pos	0.00200	1342	1279	63	55.64%
		Negative 1528		Neg	0.03700	705	31	674	
0.340	0.390	Bin 3	90%	Pos	0.00400	1368	1233	135	80.78%
		Total 3730		Neg	0.01700	1645	162	1483	
		Positive 1751	95%	Pos	0.00120	1107	1052	55	63.14%
		Negative 1979		Neg	0.03000	1248	61	1187	
0.390	1	Bin 4	90%	Pos	0.0023	619	560	59	91.43%
		Total 3513		Neg	0.00780	2593	256	2337	
		Positive 980	95%	Pos	0.00029	133	127	6	61.80%
		Negative 2533		Neg	0.02420	2038	101	1937	
Total coverage for confidence 90%				81.45%					
Total coverage for confidence 95%				56.95%					

To derive the rules we used relative importance (influence) of features evaluated during the construction of gradient boosted trees [10]-[12]. Among others, we considered features based on frequencies of all possible substrings of length at most three in the alphabet {A, C, G, T}, separately for upstream and downstream regions surrounding the motif. The most important frequency features found were the frequency of adenine in the upstream and the frequencies of adenine molecules and adenine molecule triplets in the downstream region.

Based on these findings we selected as one model feature F_A the frequency of occurrence of adenine molecules in a sequence but, as previously remarked, without the considering those within the motif.

Further investigation revealed that features based on lengths of contiguous adenine chains does have some predictive power and based on a systematic exploration in the sequence space we selected the feature C(A) that represents the maximum length of contiguous adenine chains in the region of 33 base pairs immediately following the motif on the 3' side as the feature with the best predictive power. We found that F_A is negatively and C(A) positively correlated with the probability of the enclosed motif to be a true poly(A) signal. Thus we combined these observations into a single feature $F_{VAL} = F_A^{C(A)}$ that turned out to be the most important among the considered predictive models.

For each target confidence level t ∈ {90%, 95%} we select a pair of thresholds {PosThreshold(t), NegThreshold(t)} such that the following two rules apply:

(1)If $F_{VAL} < PosTreshold(t)$ then the sample is classified as containing a true signal motif;

(2)If $F_{VAL} > NegThreshold(t)$ then the sample is classified as containing a false signal motif.

The results of this analysis are shown in Table I. The confidence levels represent the percentage of true positives or true negatives classified correctly by this rule and therefore correspond to the sensitivity and specificity of the predictive power of the rule within the rule coverage area. The coverage refers to the proportion of all samples considered in the original dataset that are covered by rules (1) and (2). We noticed however, that the value for confidence of the rules varies with values of F_A and in order to assess the predictive power of the rules above more accurately and establish the values of threshold in such a way to maximize the predictive power of the model, we increased the granularity of the analysis. For that reason we categorized samples from the original dataset into a four bins that correspond to the quartiles of distribution of F_A as we believe that quartiles represent an adequate compromise between increasing the complexity of the rule and the improvement in the rule accuracy. The thresholds in (1) and (2) are then calculated for each of the bins and the results are shown in Table I. The coverage refers to the coverage for individual bins, whereas the total coverage is summarized at the bottom of Table I. We notice that the coverage is higher for 90% confidence case than in the case for 95%, as expected. The main advantage of this predictive rule is that it is not sensitive to variant of poly (A) signal motif. In most cases, the existing prediction tools deal with only one variant or, alternatively, they represent an ensemble of separate rules built for each variant.

III. Conclusions

This decision rules derived in this work are intended to attempt to identify features relevant to polyadenilation process that are derived from the sequences surrounding the poly (A) signal motifs in human genome. It is hoped that utilization of decision rules may provide an additional insight into the genome structure. The current findings should be viewed as an initial step towards that objective. Further efforts are required to broaden the coverage by the ensemble of decision trees for polyadenilation process. In addition, similar approaches can be used for classification of other genomic signals. The reported simple rule shows good discriminating power on available data within the applicable coverage area. Since the dataset we used is smaller than the set of transcripts, it therefore represents a subset of poly (A) signals and further work should be done on extended human data as well as the genomes of higher mammalian species. This could potentially reveal conserved regulatory features. At present, we believe that the rule in the form presented here could be useful as a filter in building more complex poly (A) signal motif classifiers.

Acknowledgment

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST).

References

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

[2] E. Beaudoing, S. Freier, J. R. Wyatt, J. Claverie, and D. Gautheret, "Patterns of variant polyadenylation signal usage in human genes," *Genome Res.*, vol. 10, pp. 1001–1010, 2000.

[3] 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 Physiology*, vol. 138, pp. 1457–1468, 2005.

[4] F. Ahmed, M. Kumar, and G. Raghava, "Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies," in *Silico Biology*., vol. 9, pp. 135-48., 2009.

[5] M. N. Akhtar, S. A. Bukhari, Z. Fazal, R. Qamar, and I. A. Shahmuradov, "POLYAR, a new computer program for prediction of poly(A) sites in human sequences," *BMC Genomics*, vol. 11, p. 646, 2010.

[6] M. Legendre and D. Gautheret, "Sequence determinants in human polyadenylation site selection," *BMC Genomics*, vol. 4, p. 7, 2003.

[7] H. Liu, H. Han, J. Li, and L. Wong, "An in-silico method for prediction of polyadenyl-ation signals in human sequences," *Genome Inform.*, vol. 14, pp. 84-93, 2003.

[8] J. E. Tabaska and M. Q. Zhang, "Detection of polyadenylation signals in human DNA sequences," *Gene*, vol. 231, pp. 77-86, 1999.

[9] M. Kalkatawi, F. Rangkuti, M. Schramm, B. R. Jankovic, A. Kamau, R. Chowdhary, J. A. C Archer, and V. B. Bajic, "Dragon PolyA spotter: Predictor of poly(A) motifs within human genomic DNA sequences," *Bioinformatics*, vol. 28, no. 1, pp. 127-129, Jan. 1, 2012.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[11] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, pp. 367-378, 2002.

[12] T. Hastie, R. Tibshirani, and J. H. Friedman, *Elements of Statistical Learning: Data Mining, Inference and Prediction* 2nd ed., Springer-Verlag, New York, 2009.

Hassan Abou Eisha is a PHD student with the CEMSE Division at KAUST. He graduated from the German University in Cairo in the field of Computer Sciences and Engineering in 2010. He received his Master of Science in Computer Sciences from KAUST in 2011. His main research interests are concerned with computational complexity, discrete optimization and graph theory.