Journal of Theoretical Biology 265 (2010) 287–296

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A classification-based prediction model of messenger RNA polyadenylation sites

Guoli Ji [a,*], Xiaohui Wu [a], Yingjia Shen [b], Jiangyin Huang [a], Qingshun Quinn Li [b,**]

[a] Department of Automation, Xiamen University, Xiamen 361000, China
[b] Department of Botany, Miami University, Oxford, Ohio 45056, USA

ARTICLE INFO

Article history:
Received 13 December 2009
Received in revised form
21 March 2010
Accepted 13 May 2010
Available online 26 May 2010

Keywords:
Arabidopsis
Classification-based modeling
Genome annotation
Polyadenylation
Predictive modeling

ABSTRACT

Messenger RNA polyadenylation is one of the essential processing steps during eukaryotic gene expression. The site of polyadenylation [(poly(A) site] marks the end of a transcript, which is also the end of a gene. A computation program that is able to recognize poly(A) sites would not only prove useful for genome annotation in finding genes ends, but also for predicting alternative poly(A) sites. Features that define the poly(A) sites can now be extracted from the poly(A) site datasets to build such predictive models. Using methods, including *K*-gram pattern, *Z*-curve, position-specific scoring matrix and first-order inhomogeneous Markov sub-model, numerous features were generated and placed in an original feature space. To select the most useful features, attribute selection algorithms, such as information gain and entropy, were employed. A training model was then built based on the Bayesian network to determine a subset of the optimal features. Test models corresponding to the training models were built to predict poly(A) sites in Arabidopsis and rice. Thus, a prediction model, termed Poly(A) site classifier, or PAC, was constructed. The uniqueness of the model lies in its structure in that each sub-model can be replaced or expanded, while feature generation, selection and classification are all independent processes. Its modular design makes it easily adaptable to different species or datasets. The algorithm's high specificity and sensitivity were demonstrated by testing several datasets and, at the best combinations, they both reached 95%. The software package may be used for genome annotation and optimizing transgene structure.

1. Introduction

After transcription from genomic DNA, precursor messenger RNA (mRNA) needs to be processed before becoming functional in eukaryotic cells. One of the essential processing event is the addition of a polyadenine [poly(A)] track to a terminal nucleotide of the 3′-untranslated region (3′-UTR). The location of this terminal nucleotide, which is exposed after endonuclease cleavage, is known as a poly(A) site. Thus, a poly(A) site marks the end of the transcribed mature mRNA, and, as such, it can be used to find and annotate the end of a gene. Identification of poly(A) sites also facilitates the search for genes that undergo alternative polyadenylation, a significant mode of gene expression regulation that is increasingly observed in animal and plant genes (Delaney et al., 2006; Lutz, 2008; Quesada et al., 2005; Xing et al., 2008; Zhang et al., 2005). Moreover, since there are instances where foreign transgenes may carry unwanted poly(A) sites, the use of such poly(A) sites during transgene expression may destroy the function of the transgene in the target organisms (Diehn et al., 1998). Thus, identification and elimination of these cryptic poly(A) sites would be of interest in biotechnological applications.

The location of a poly(A) site for a gene is mostly predetermined by the so-called polyadenylation signals. Traditionally, poly(A) sites are identified by examining the expressed sequence tags (ESTs) which are reverse transcribed from mature mRNA. Since the poly(A) tail is added post-transcriptionally, alignment of ESTs to their respective genomic sequences will reveal the location of poly(A) sites. Indeed, there are number of datasets with collections of poly(A) sites (Graber et al., 1999; Loke et al., 2005; Shen et al., 2008a, b; Zhang et al., 2005). Further analysis of these datasets have elucidated the poly(A) signals that determine the poly(A) site locations at the genome level. Such information about the poly(A) signals, particularly those from Arabidopsis (Loke et al., 2005) and rice (Shen et al., 2008a), is the foundation of our work which involves the construction of predictive models for the systematic prediction of plant poly(A) sites.

The complexity of poly(A) sites is demonstrated by the fact that poly(A) sites can be located in a short region of the 3'-UTR (Li and Hunt, 1997; Loke et al., 2005; Shen et al., 2008a).

* Corresponding author. Tel.: +86 5922181049.
E-mail addresses: glji@xmu.edu.cn (G. Ji), liq@muohio.edu (Q. Quinn Li).
** Corresponding author. Tel.: +1 513 529 4256.

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.05.015

Furthermore, plant poly(A) signals possess little conservation. These properties, coupled with a limited knowledge of numerical prediction and its application to plant polyadenylation, make it hard to predict exact poly(A) sites using computational methods. Therefore, to overcome these inherent obstacles to accurate plant gene annotation and transgene design, we first developed a new, modulated algorithm that is composed of three independent components (feature generation, selection and classification) and reaches maximum flexibility in its adaptation to new features. Then, we employed these feature-selection algorithms to select a relative best feature space that includes only effective features. Finally, we built a recognition model, termed Poly(A) site Classifier, or PAC, to effectively predict poly(A) sites.

2. Generation of signal feature space and poly(A) classifier

2.1. The datasets

A dataset of Arabidopsis 3′-UTR of 8,160 poly(A) site entries, which we termed the 8 K dataset, was described previously (Loke et al., 2005). The poly(A) sites were authenticated by comparing ESTs to the genomic sequences. After mapping ESTs to the genome, coordinates of the poly(A) sites in the genomic sequences were defined as the last nucleotide that matched to the genome sequences. Then, the genomic sequences of 400 nucleotides (nt) in length, including 301 nt upstream to 99 nt downstream of the poly(A) site, were extracted. Hence, the known poly(A) site for each sequence was located at the 301st nt (counting from left to right). The reason for the length of the sequences was set to 400 nts was based on previous analysis (Loke et al., 2005) where all poly(A) signals to determine the poly(A) sites were included. In these operations, we used DNA sequence in place of RNA, but there was no impact on modeling. The rice dataset, which was generated in the same manner as Arabidopsis, has been described elsewhere (Shen et al., 2008a). It contains 55,000 annotated poly(A) sites, and we termed it the 55 K dataset. For the 8 K dataset of Arabidopsis, there are 6059 unique genes represented in the dataset. Among them, 78.6% of them have only one poly(A) site, 15.7% have two sites, 3.5% have three sites, and 2.2% have four or more sites. The majority of the poly(A) sites (96% of 8160) were located in the 3′-UTR as expected, while 1.8% located in the introns (0.3%), coding sequences (0.4%) and 5′-UTR (1.1%), and 2.2% located in the intergenic regions. For rice 55 K dataset, a total of 16,911 unique genes were represented (Shen et al., 2008a); out of which, 49.0% show only one poly(A) site, 24.1% have two sites, 13.2% have three sites, and 13.6 have 4 or more sites. Among the 55 K poly(A) sites, 86.9% were found in the 3′-UTR, 1.9% located in the introns (0.93%), coding sequences (0.45%), and 5′-UTR (0.54%), while 11.1% were located in the intergenic regions (Shen et al., 2008a).

To make the model suitable for variable poly(A) site recognition, the training datasets were derived from several different kinds of sequences. A training dataset consisting of 487 (an arbitrary number) positive sequences was randomly extracted from the 8 K sequences (Ji et al., 2007a; Loke et al., 2005). A negative training dataset consists of the following: 100 randomly generated sequences that preserved the trinucleotide distributions in 3′-UTR using the Markov Chain (hence, it is called MC); 100 5′-UTRs,100 introns and 100 coding sequences of Arabidopsis were extracted from the databases of the Arabidopsis Information Resources (TAIR) as described previously (Ji et al., 2007a). The sequences in the training dataset were trimmed into lengths of 162 nt each, the size of the scanning window. This window size was based on the profile of Arabidopsis nucleotide sequence distribution and polyadenylation signals around the poly(A) sites, as we described before (Ji et al., 2007a; Loke et al., 2005), from which it is reasonable to assume that each poly(A) site is only correlated with the poly(A) signals upstream and downstream of it. In this paper, a window sequence is defined as the sequence containing 131 nt upstream and 31 nt downstream of a poly(A) site (Loke et al., 2005); thus internal processing sliding window sequence is 162 nt in length.

As another control, a negative training dataset consisting of 1100 sequences was randomly chosen from the 8 K dataset, each extracted from sequences at least 10 nt beyond both sides of the poly(A) site. Longer sequences (>162 nt) were cut into shorter 162 nt sequences for processing, and then the outputs were joined together to reflect the outcome of the original sequences.

In order to evaluate the PAC prediction results, as described below, the test datasets (different from the training datasets) were made up of 35 long sequences [produced by merging some of the 400 nt sequences from the dataset that come from the same gene so to make longer sequences for Sn calculations; each of the sequences thus would have multiple poly(A) sites] with 154 known multi-poly(A) sites and 100 each of the following types of sequences (each 400 nt in length): randomly generated 3′-UTRs but preserved the original trinucleotide distribution, 5′-UTRs, introns and coding sequences of Arabidopsis, respectively.

2.2. Generation of signal feature space

When using a classification algorithm for predicting the poly(A) sites, the nucleotide sequence needs to be converted into numeric format. Consequently, the features of poly(A) signals around the cleavage sites were extracted based on the profile of nucleotide sequence distribution around the poly(A) sites in both Arabidopsis and rice. To deal with complicated biological problems by classifier models, single feature is not nearly enough, so ensemble features are being increasingly used to construct classifier (Chou and Shen, 2007; Frey et al., 2007; Kedarisetti et al., 2006; Shen and Chou, 2006, 2009). In practical applications, particularly in developing high throughput tools for predicting various important attributes for biomacromolecules, many different descriptors to represent biological sequence samples have been developed and widely used, such as those by means of cellular automata image (Lin et al., 2009; Xiao et al., 2006, 2009, 2008b), those by complexity measure factor (Xiao et al., 2006, 2005), and those by grey dynamic model (Lin et al., 2009; Xiao et al., 2008a), as well as many other feature representation methods (Chou, 2009). Here, five feature representation methods were adopted to describe the makeup of nucleotide sequences. These methods were chosen to confirm whether each one could generate unique features from different training datasets. Finally, the numerical vector was used as the input of the classification algorithm. The distribution of features in different areas of a window sequence is shown in Fig. 1.

K-gram nucleotide sequence pattern: Given a *K*-gram (a subsequence of *K* nucleotides) and a scanned region of length *L*, the relative probability of this *K*-gram can be obtained by scanning from the first to the last position ($L-K+1$) of the scanning region (Liu et al., 2003). Using four mono-nucleotides and sixteen di-nucleotides, we scanned upstream ($-1\sim-130$ nt) and downstream ($+1\sim+32$ nt, where the poly(A) site is defined at -1 position) of the poly(A) sites to get 40 distinct mono-nucleotide and di-nucleotide probabilities (each *K*-gram corresponding to 2 probabilities) as a part of the initial feature space.

Z-Curve: the *Z*-curve (Zhang and Wang, 2000) is a three-dimensional space curve reconstituting each unique DNA/RNA sequence; thus, each sequence can be represented as a *Z*-curve, and the sequence and *Z*-curve can be reconstituted from each other.