RECOGNITION OF POLYADENYLATION SITES FROM ARABIDOPSIS GENOMIC SEQUENCES

CHUAN HOCK KOH kohchuan@comp.nus.edu.sg
LIMSOON WONG wongls@comp.nus.ed.sg

School of Computing, National University of Singapore
COM1, Law Link, Singapore 117590

A polyadenine tail is found at the 3' end of nearly every fully processed eukaryotic mRNA and has been suggested to influence virtually all aspects of mRNA metabolism. The ability to predict polyadenylation site will allow us to define gene boundaries, predict number of genes present in a particular gene locus and perhaps better understand mRNA metabolism. To this end, we built an arabidopsis polyadenylation prediction model. The prediction model uses a machine learning method which consists of four sequential steps: feature generation, feature selection, feature integration and cascade classifier. We have tested our model on public datasets and achieved more than 97% sensitivity and specificity. We have also directly compared with another arabidopsis prediction model, PASS 1.0, and have achieved better results.

Keywords: arabidopsis, machine learning, polyadenylation site

1. Introduction

Polyadenylation is a post-transcriptional process. The process basically cleaves and adds about 200 adenosine residues to the pre-mRNA 3' end. The site where the pre-mRNA is cleaved is known as the polyadenylation site. The selection of polyadenylation sites are determined by polyadenylation signals or cis-elements in the pre-mRNA. In humans, AAUAAA is a highly conserved polyadenylation signal. However, no highly conserved polyadenylation signal has been identified in arabidopsis. In this respect, the prediction of polyadenylation site is therefore more difficult.

The polyadenine tail has been shown to boost translation, protects the 3' end of mRNA from exonucleases and is needed for the mRNA nuclear-to-cytoplasmic export. This process has also been found to be tightly coupled with splicing and transcription termination. Thus, it is an essential processing event and an integral part of gene expression [5].

Therefore, the ability to predict the polyadenylation site potentially allows us to better understand the process and also to be able to better segment genes. To this end, we developed an arabidopsis polyadenylation prediction model.

Although proteins involved in polyadenylation processes appears to be conserved among human and arabidopsis, their polyadenylation signals differ widely in terms of their locations with respect to polyadenylation site and sequence content [5]. In humans, AAUAAA (or its one-base variant) is a highly conserved polyadenylation signal and is found in 87.1% of the observed sites [1]. In plants, there are no highly conserved

polyadenylation signals. AAUAAA is the most frequently occurring polyadenylation signal, yet it is found in only about 10% of arabidopsis genes [5].

Nonetheless, studies have shown that arabidopsis polyadenylation signals are composed of three major groups: far upstream elements (FUE), near upstream elements (NUE) and cleavage elements (CE). The far upstream elements span a region of 60-130 nucleotides, resides at a location 18 to 22 nucleotides upstream of the cleavage site and has a high U content. The near upstream elements spans a region of 6-10 nucleotides, resides at a location 12 nucleotides upstream of the cleavage site and has a high A content. The cleavage elements span a region of about 27 nucleotides, resides from a position which is 12 nucleotides upstream to about 15 nucleotides downstream of the cleavage site and it consists of the well conserved YA(CA or TA) just before the cleavage site and U-rich elements flanking both sides of the cleavage site [5]. For a schematic representation of polyadenylation site in arabidopsis mRNA 3' ends, see Figure 1.

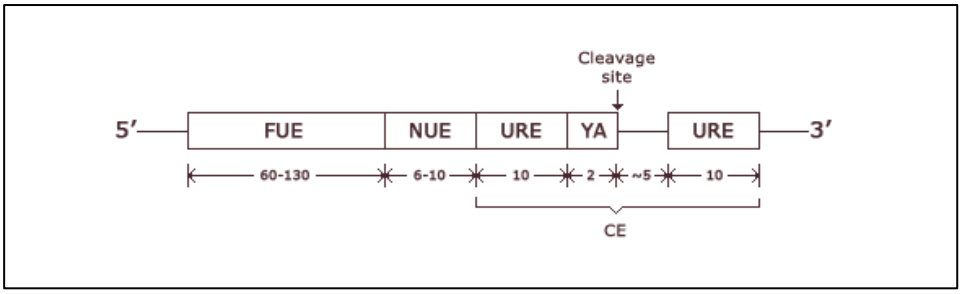

Figure 1. A schematic representation of polyadenylation site in arabidopsis mRNA 3' ends.

Currently, there exists another prediction program for arabidopsis named polyadenylation site sleuth or PASS [3]. PASS was developed based on Generalized Hidden Markov Model using polyadenylation signals identified by [5]. We compared against PASS 1.0 using the same datasets and achieved better results in most cases.

For our arabidopsis polyadenylation prediction, we built a model based on the machine learning methodology described in [4]. However, we added an additional step to that methodology - a cascade classifier. This additional step is implemented due to the fact that there are no highly conserved polyadenylation signals in Arabidopsis; and this step has been shown to increase both sensitivity and specificity.

2. Tools, Datasets, and Methods

2.1 *Tools*

In our prediction model, certain steps were implemented using Waikato Environment for Knowledge Analysis (WEKA). WEKA is a free machine learning software package written in Java and developed at University of Waikato [8]. One particular learning

scheme that we have employed from WEKA is SMO. SMO is the WEKA implementation of support vector machine [7] using John Platt's sequential minimal optimization algorithm [6].

2.2 *Datasets*

Datasets used for this project came from two main sources.
1) Datasets derived by Hao Han

- Dataset A (used to set parameters):
    - 804 (+ve) sequences with EST-supported polyadenylation sites, derived based on ATPACDB and has confidence "high" or "very high". (http://harlequin.jax.org/atpacdb/confidence.php)
    - 9742 (-ve) coding sequences that were extracted from ENSEMBL database, arabidopsis section
    - Sequences in Dataset A are of length 400. For each of the 804 (+ve) sequences, the EST-supported polyadenylation is at location 201.

2) Datasets provided by Qingshun Li; please refer to [3] for more information on how the following datasets were derived

- Dataset B (used for SMO1 training):
    - 2640 (+ve) sequences with EST-supported polyadenylation sites
    - 900 (-ve) coding sequences
    - 476 (-ve) 5'UTR sequences
    - 954 (-ve) intronic sequences
- Dataset C (used for SMO2 training):
    - 1500 (+ve) sequences with EST-supported polyadenylation sites
    - 100 (-ve) coding sequences
    - 100 (-ve) 5'UTR sequences
    - 100 (-ve) intronic sequences
- Dataset D (used for SMOA and SMO2 testing):
    - 2069 (+ve) sequences with EST-supported polyadenylation sites
    - 501 (-ve) coding sequences
    - 288 (-ve) 5'UTR sequences
    - 527 (-ve) intronic sequences
- Dataset E (used for SMOA training):
    - 4140 (+ve) sequences with EST-supported polyadenylation sites
    - 1000 (-ve) coding sequences
    - 576 (-ve) 5'UTR sequences
    - 1054 (-ve) intronic sequences
- Each sequence in Dataset B, C, D and E is of length 400. Each (+ve) sequences has the EST-supported polyadenylation sites at location 301. Dataset E is formed by combining Dataset B and Dataset C. Each sequence in Dataset B, C and D underwent pair-wise global alignment against every other sequence. If

any two sequences have more than 70% similarity, one of them is discarded. (i.e., if we randomly pick a sequence from Dataset B, we will not find any other sequence in Dataset B, C or D with more than 70% similarity). This is to minimize biasness due to similarity of sequences.

2.3 *Methods*

2.3.1 *Architecture of the polyadenylation prediction system*

Our polyadenylation prediction system is a cascade of two layers of classifiers. In the first layer, a classifier SMO1 is used to score positions (-40/+40) at every nucleotide and therefore, there is a total of 81 SMO1 scores relative to a candidate site. In the second layer, a classifier SMO2 takes these 81 SMO1 scores to decide if the candidate site is a polyadenylation site.

We follow the "feature generation, feature selection, feature integration" methodology [4] in developing our prediction system, and in particular, the first-layer classifier SMO1.

The architecture of our polyadenylation prediction system and an overview of the steps involved are depicted in Figure 2.

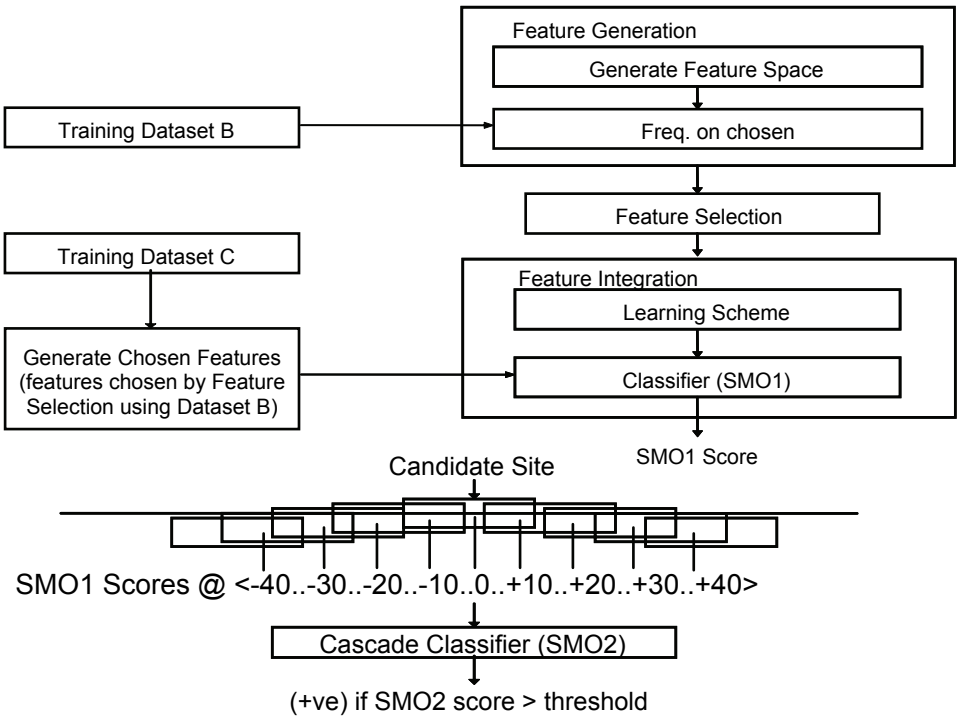

Figure 2. Architecture of our polyadenylation prediction system.

2.3.2 *Training Phase*

1) Feature Generation

The candidate features that we considered for SMO1 are 1-gram (A, C, G, U), 2-gram (AA, AC, AG, AU,…, GU, UU), 3-gram (AAA, AAC, AAG, AAU, …, UGU, UUU), 4U/1N (NUUUU, UNUUU, UUNUU, UUUNU, UUUUN), 4A/1N (NAAAA, ANAAA, AANAA, AAANA, AAAAN) and G/U*7 (A stretch of G or U for 7 bp). We consider these features separately in 3 different windows relative to the candidate site, (-110/+5), (-35/+15) and (-50/+30); and we calculate the frequency of the features in these 3 different windows. A total of 261 candidate features were generated.

We chose these features and windows by referencing to biological literature [1, 2 and 5] and findings from analyzing Dataset A.

2) Feature Selection

The feature selection step is done by using WEKA supervised attribute filter with Attribute Evaluator set to "ChiSquaredAttributeEval" and Search Method set to "Ranker". Statistics of these candidate features are then computed based on Dataset B for SMO1. Features that have chi-square statistics greater than the threshold value of 0 are chosen to form the feature vector.

Out of the 261 candidate features generated, 228 have chi-square value exceeding 0 and were selected. Of them, the minimum observed chi-square value is 11.

3) Feature Integration

The first-layer classifier SMO1 is then trained using the SMO support vector machine learning scheme in WEKA and the selected features from Step 2. The training data used in this step is from Dataset B.

4) Cascade Classifier (SMO2)

The second layer classifier SMO2 is then trained using the SMO support vector machine learning scheme in WEKA. The feature vector for this step is the 81 scores output by SMO1 at positions (-40/+40) relative to a candidate site. The training data for this step is from Dataset C.

2.4 *Prediction Phase*

Although in Figure 1, "YA" is said to be well conserved at positions just before the polyadenylation sites, it is found in less than 41% of the sequences in Dataset A. Therefore, when given a sequence to predict the existence and location of polyadenylation sites, we consider every location to be a possible candidate site instead of only "YA" positions. Hence, the classifier SMO1 is first deployed to get a score at every nucleotide for a given sequence. Cascade classifier SMO2 then makes use of the SMO1 scores at positions (-40/+40) relative to a candidate site to carry out prediction for polyadenylation sites.

However, in order to make predictions on a particular nucleotide, SMO1 needs to know that particular nucleotide's (-110/+30) composition. Therefore, for SMO2 to make a prediction, it needs to know the composition of (-150/+70). Hence, given a sequence of DNA, our prediction model is not able to make prediction on the first 150 and last 70 nucleotides of the sequence. Likewise, PASS 1.0 is also unable to make predictions on the first 149 and last 10 nucleotides. PASS 1.0 assumes these locations to be unlikely to be a polyadenylation site and sets the scores for these locations to be 0. Our prediction model also assumes these locations to be unlikely to be a polyadenylation site and sets their scores to 0. Therefore, the users of our model and/or PASS 1.0 are advised to provide sequences that are of sufficient length. (i.e., having upstream and downstream of a sequence to go beyond 3'UTR).

3. Results

In order to show improvements given by using the cascade classifier step, we also trained another classifier SMOA with Dataset E using the same steps for SMO1. Using Dataset E for SMOA ensures that SMOA and SMO2 uses equal amount of training data.

We then used Dataset D to evaluate the performance of our polyadenylation prediction system (SMO2), SMOA and PASS 1.0. We considered the following measures:

$$\text{Sensitivity (SN)} = TP / (TP + FN)$$
$$\text{Specificity (SP)} = TN / (TN + FP)$$

where TP (True Positive) is the total number of EST-supported polyadenylation sites that are identified or correctly predicted in the (+ve) sequences. FN (False Negative) is the total number of EST-supported polyadenylation sites that are not identified or predicted correctly in the (+ve) sequences. TN (True Negative) is the total number of sites with score ≤ threshold in the (-ve) sequences. FP (False Positive) is the total number of sites with score > threshold in the (-ve) sequences.

The sensitivity and specificity of SMOA, SMO2 and PASS 1.0 running on the test dataset can be observed in Figures 3, 4 and 5 respectively. The equal-error-rates values (i.e., the points where sensitivity = specificity) achieved by SMOA, SMO2 and PASS with SN_0, SN_10 and SN_30 are laid out in tabular form on Tables 1, 2 and 3.

SN_0 means the predicted polyadenylation site is exactly the same as the EST-supported polyadenylation site. SN_10 means the EST-supported polyadenylation site is within 10 nucleotides of the predicted polyadenylation site. SN_30 means the EST-supported polyadenylation site is within 30 nucleotides of the predicted polyadenylation site. SP_CDS means the specificity achieved by running the classifier on coding sequences. SP_5UTR means the specificity achieved by running the classifier on 5'UTR sequences. SP_Intron means the specificity achieved by running the classifier on intronic sequences.

For Figure 5, the specificity of PASS 1.0 differs from what was reported in [3]. This is because the specificity calculated in [3] includes positions even where PASS 1.0 was

unable to make a prediction as stated in Section 2.4. By doing so, it almost always recognizes those positions as TN, which will inevitably increase its specificity, whereas the calculation of specificity in this paper does not make use of those positions, hence resulting in a lower specificity.

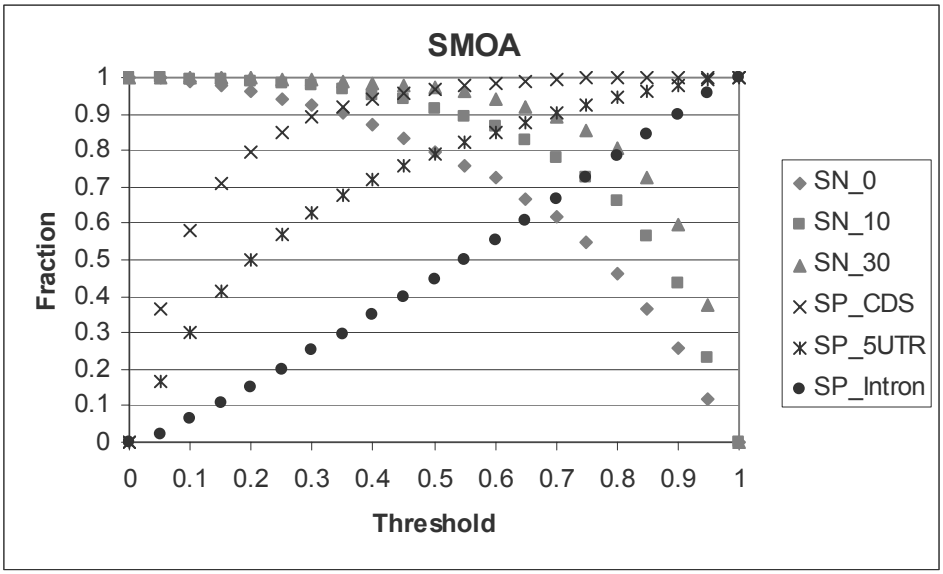

Figure 3. The prediction performance of SMOA.

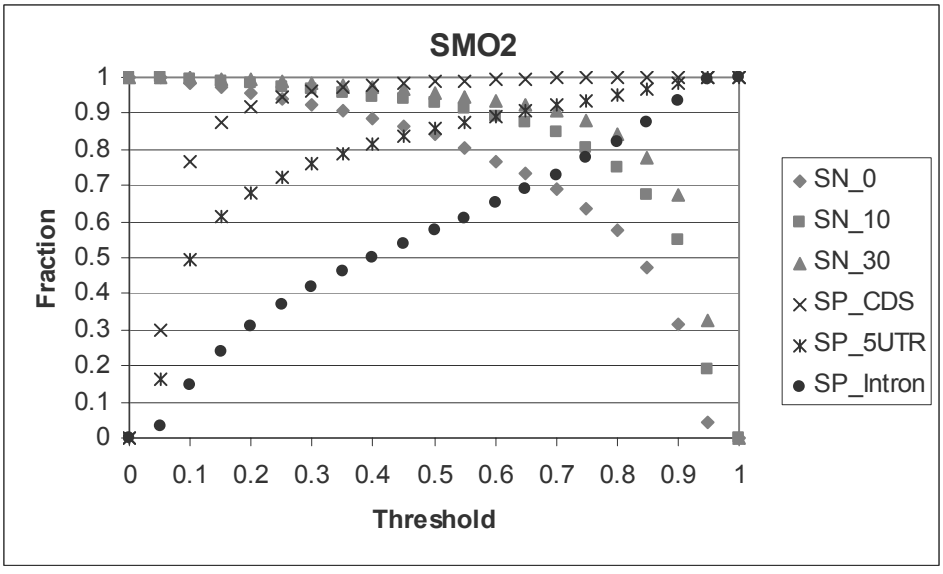

Figure 4. The prediction performance of SMO2.

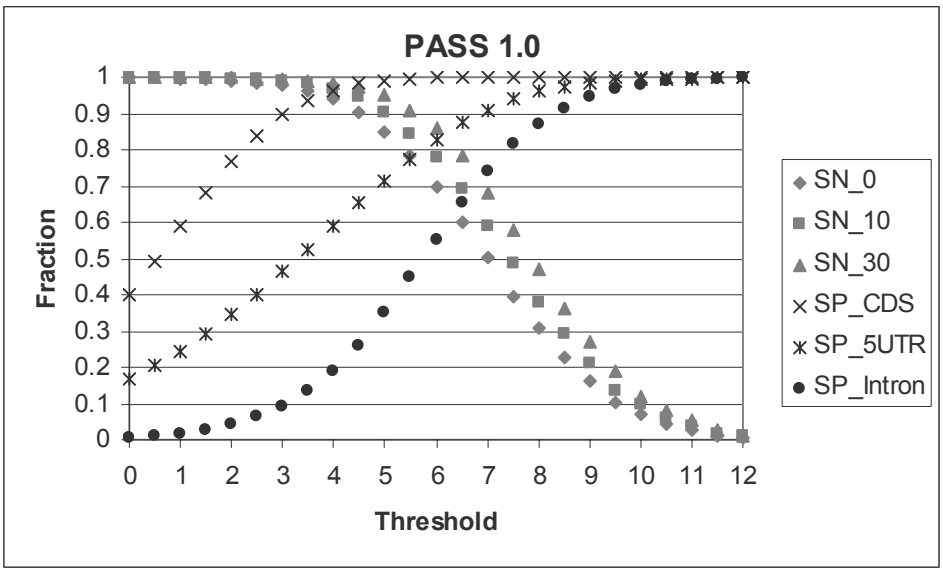

Figure 5. The prediction performance of PASS 1.0.

Table 1. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN_0.

SN_0	**SMO A**		**SMO 2**		**PASS 1.0**	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	91.1%	0.33	94.3%	0.24	95.3%	3.76
5'UTR	79.3%	0.50	84.9%	0.48	77.7%	5.53
Intron	63.9%	0.68	71.1%	0.68	62.8%	6.36

Table 2. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN_10.

SN_10	**SMO A**		**SMO 2**		**PASS 1.0**	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	94.8%	0.42	96.5%	0.31	96.5%	4.02
5'UTR	85.8%	0.61	89.2%	0.60	80.7%	5.81
Intron	72.5%	0.75	78.8%	0.76	67.7%	6.62

Table 3. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN_30.

SN_30	**SMO A**		**SMO 2**		**PASS 1.0**	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	97.1%	0.50	97.5%	0.37	97.5%	4.29
5'UTR	89.8%	0.69	91.5%	0.67	84.0%	6.13
Intron	79.2%	0.81	83.0%	0.81	71.7%	6.85

4. Discussions

From the results, it is clear that introducing the cascade classifier (SMO2) step on top of the "feature generation, feature selection, feature integration" methodology helps improve both sensitivity and specificity (Shown by comparison between SMOA and SMO2). Our prediction model (SMO2) is also able to achieve a significant 7 – 11% higher sensitivity and specificity against PASS 1.0 with control sequences from 5'UTR and introns, while maintaining similar sensitivity and specificity with control sequences from the coding region.

When different control sequences are used, a drastic difference in the levels of sensitivity and precision is observed. With SN_0, coding sequences gives a sensitivity and specificity of 94.3% while introns sequences only achieved 71.1%; see Table 1. There could be two main reasons for this:

Firstly, recall that the parameter of our arabidopsis prediction model is set by using Dataset A derived by Hao Han. In that dataset, the control sequences used are from coding regions. This could have caused our arabidopsis prediction model to be biased towards coding sequences. Introducing 5'UTR and intronic sequences into the dataset used for setting the parameters could potentially increase the accuracy achieved on 5'UTR and intron.

Secondly, features used in our arabidopsis prediction model are compositional features. It is known that the intronic sequences have high A and U composition. This is a characteristic shared by the (+ve) sequences with EST supported polyadenylation site. This could explain why our arabidopsis prediction model has more difficulty in separating them. Therefore, introducing positional specific features or any features that are not compositional could potentially enhance the prediction model.

3. Conclusions

Our arabidopsis prediction model was built based on a machine learning methodology as described in [4]. It basically consists of 3 steps: 1) feature generation 2) feature selection 3) feature integration. An extra step (cascade classifiers) has been added for our model which helped increase sensitivity and specificity. In the above few steps, feature generation is the most crucial and difficult step. Given the "correct" set of features to generate, any method used for feature selection, feature integration and cascade classifier would yield high accuracy. However, to find the "correct" set of features to generate is similar to the problem of motif finding given a set of sequences, which is known to be NP-hard. Our approach is to use biological understanding of polyadenylation mechanism to decide on the set of features to generate. This approach has been shown to yield reasonably high accuracy in our prediction model.

Our arabidopsis prediction model has outperformed PASS 1.0 by a 7 – 11% margin on a validation dataset given by Qingshun Li (inventor of PASS 1.0). On the validation dataset, sensitivity and specificity for our model ranges from 71.1% to 94.3% with SN_0. As discussed previously, it is possible that the prediction model that we have developed

for arabidopsis is biased towards coding sequences due to the parameter setting process. Therefore, one possible way to better improve its accuracy for other control sequences would be to redo the parameter setting procedure using other control sequences. Also, including more features that are not compositional could help better distinguish (+ve) sequences from (-ve) intronic sequences as they are compositionally similar.

As polyadenylation takes place just before the end of a transcription, the ability to accurately predict a polyadenylation site would be useful in predicting the ends of transcripts and also terminal exons. To this end, our prediction model has achieved reasonable accuracy and would certainly be useful for better gene annotations.

The datasets and source files used for this project are available at http://www.comp.nus.edu.sg/~wongls/projects/dnafeatures/giw07-supplement/

Acknowledgments

We thank Huiqing Liu and Hao Han for providing Dataset A and for their involvement in helpful discussions. We are also grateful to Qingshun Li for providing Dataset B, C, D and E.

References

[1] Emmanuel Beaudoing, Susan Freier, Jacqueline R. Wyatt, Jean-Michel Claverie and Daniel Gautheret, (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research*, Vol. 10, No. 7, July, 2000, pp. 1001-1010.

[2] Diana F. Colgan and James L. Manley, (1997). Mechanism and regulation of mRNA polyadenylation. *Genes & Dev*, Vol. 11, No. 21, November 1, 1997, pp. 2755-2766.

[3] Guoli Ji, Jianti Zheng, Yingjia Shen, Xiaohui Wu, Ronghan Jiang, Yun Lin, Johnny C Loke, Kimberly M Davis, Greg J Reese and Qingshun Quinn Li, (2007). Predictive modeling of plant messenger RNA polyadenylation sites. *BMC Bioinformatics*, Vol. 8, No. 43, February 7, 2007.

[4] Huiqing Liu and Limsoon Wong, (2003). Data Mining Tools for Biological Sequences. *Journal of Bioinformatics and Computational Biology*, Vol. 1, No. 1, April 7, 2003, pp. 139-167.

[5] Johnny C. Loke, Eric A. Stahlberg, David G. Strenski, Brian J. Haas, Paul Chris Wood and Qingshun Quinn Li, (2005). Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures. *Plant Physiology*, Vol. 138, July 2005, pp. 1457-1468.

[6] J. Platt, (1999). Fast Training of support vector machines using sequential minimal optimization. B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods --- Support Vector Learning*, pages 185-208, Cambridge, MA, 1999. MIT Press.

[7] V.N. Vapnik, (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.

[8] Ian H. Witten and Eibe Frank (2005). "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.