A Generic System For Genomic Feature Recognition

National University of Singapore

Student: Koh Chuan Hock and Supervisor: Professor Wong Limsoon

BACKGROUND

Polyadenylation is a post-transcriptional process. The process basically cleaves and adds about 200 adenosine residues to pre-mRNA 3' end. The site where the pre-mRNA is cleaved is known as the polyadenylation site and the selection of polyadenylation sites are determined by polyadenylation signals or cis-elements in the pre-mRNA.

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. The 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 (Loke, Stahlberg, Strenski, Haas, Wood and Li, 2005).

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. Hence, my research is to develop prediction models for polyadenylation sites with focus in arabidopsis sequences. To this end, I had developed two prediction models: for human and arabidopsis.

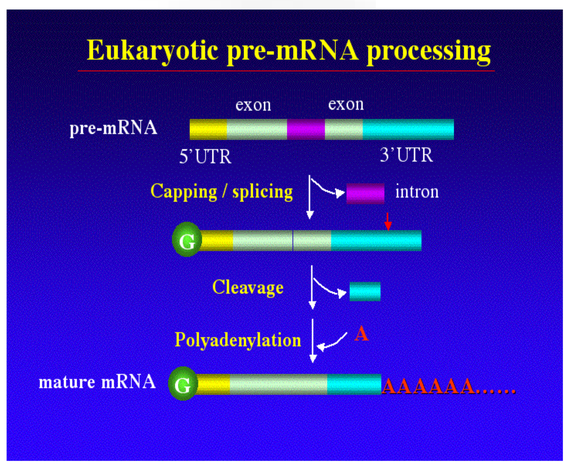

Image Credit: www.polya.org

METHODS

- Human Prediction Model: Developed based on the machine learning method described in (Liu and Wong, 2003b).
- Arabidopsis Prediction Model: Also developed based on the machine learning method described in (Liu et al., 2003b) with an additional step, cascade classifier.

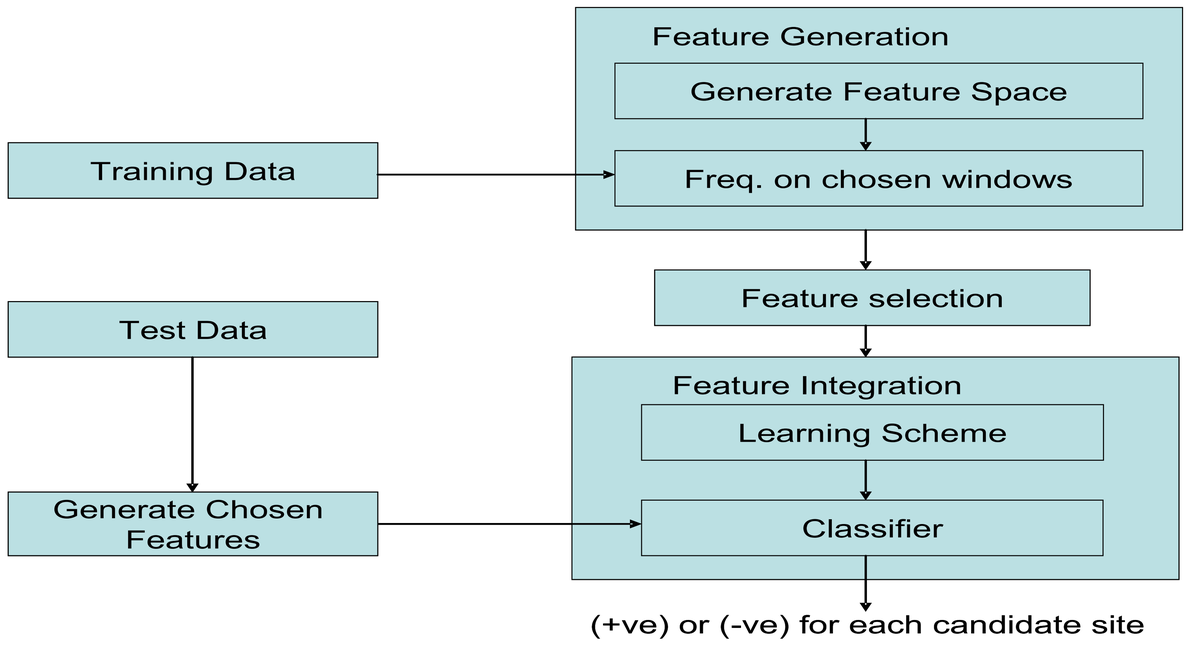

Figure 1. Human Prediction Model

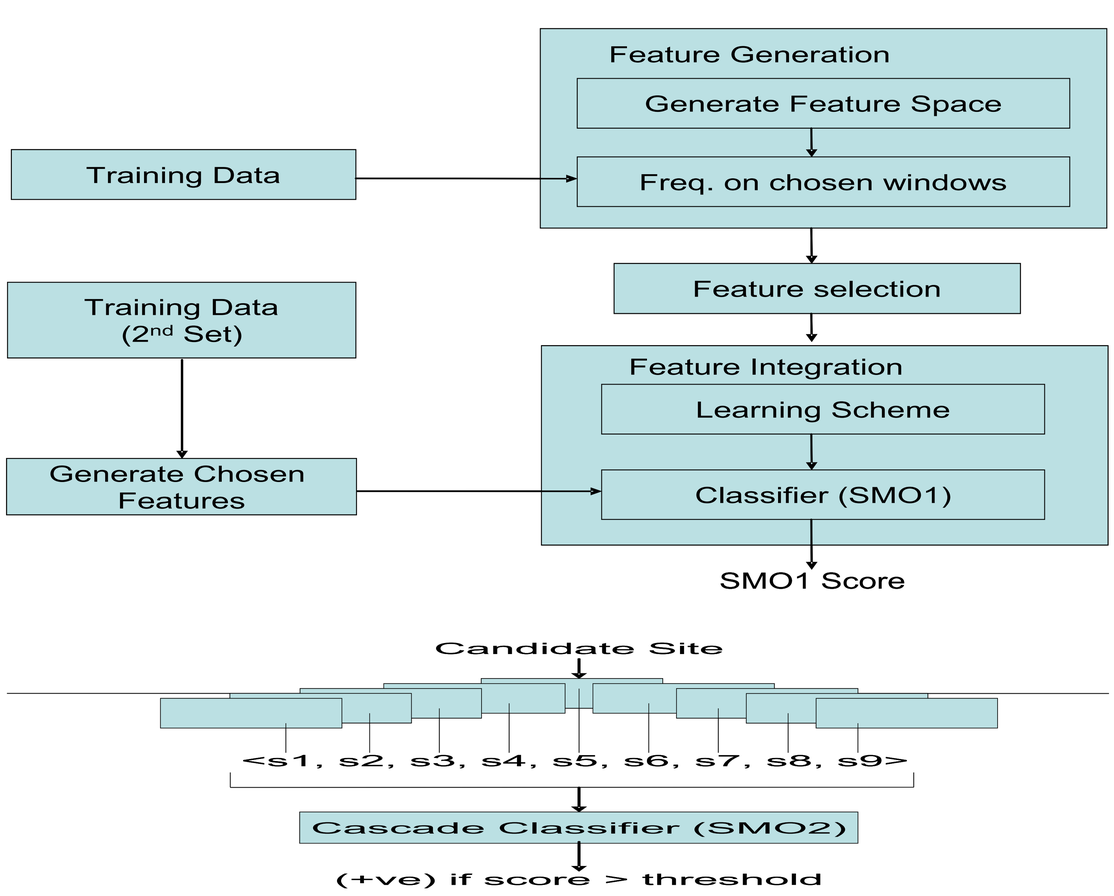

Figure 2. Arabidopsis Prediction Model

RESULTS

Human

Feature Selection		Learning Scheme			
		J48		SMO	
Chi-square		Sensitivity (%)	Precision (%)	Sensitivity (%)	Precision (%)
Liu et al.	+ve	71.1	71.9	82.5	80.0
	-ve	73.8	73.1	80.6	83.1
My approach	+ve	71.8 (+0.7)	72.1 (+0.2)	83.6 (+1.1)	82.4 (+2.4)
	-ve	74.0 (+0.2)	73.6 (+0.5)	83.2 (+2.6)	84.4 (+1.3)
CFS					
Liu et al.	+ve	73.1	72.7	82.9	79.7
	-ve	74.2	74.6	80.2	83.3
My approach	+ve	72.6 (-0.5)	74.0 (+1.3)	81.9 (-1.0)	80.7 (+1.0)
	-ve	76.1 (+1.9)	74.7 (-0.1)	81.6 (+1.4)	82.7 (-0.5)

Table 1. 10 fold cross-validation on datasets used by (Liu et al., 2003a)

Arabidopsis

Figure 3. Relationship of Sensitivity, Precision and Threshold

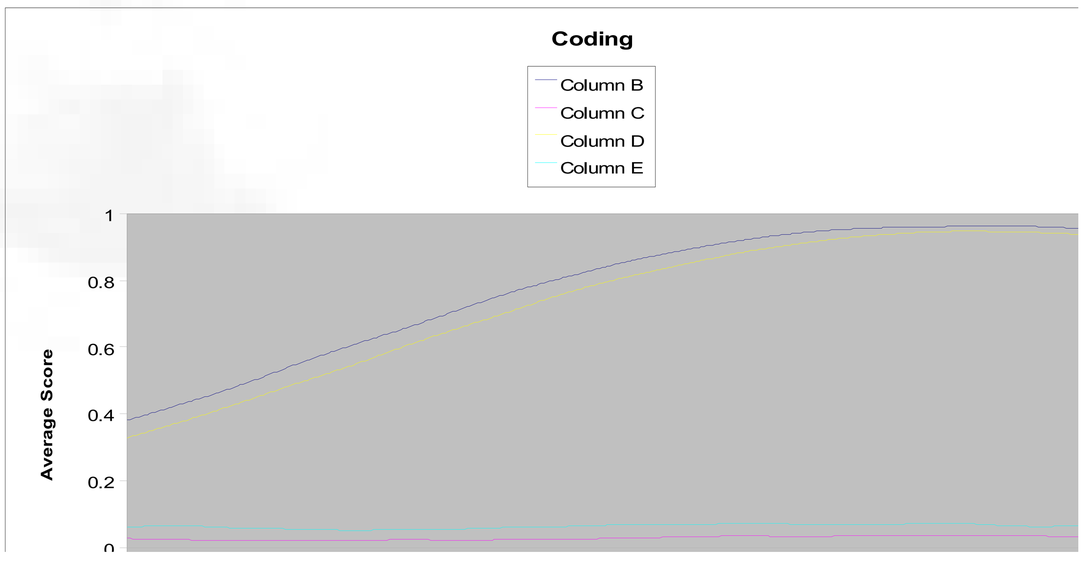

Figure 4. The average prediction scores of SMO1 and SMO2

DISCUSSION

- Human Prediction Model: Both (Liu, Han, Li and Wong, 2003a) and my human prediction model are based on the same methodology as described in (Liu et al., 2003b). The distinct difference between our model is in the feature generation step where I had incorporated biological understanding of human polyadenylation mechanism. Indeed, with this implementation, my human prediction model is able to produce higher levels of accuracy when compared to (Liu et al., 2003a); see Table1.
- Arabidopsis Prediction Model: When using coding sequences as control, my model can reach sensitivity and precision of above 96% at a threshold value of 0.95 (see Figure 3). The additional step, which is an introduction of the cascade classifier to my arabidopsis model, has been shown to help better differentiate true polyadenylation sites from pseudo ones (see figure 4).

CONCLUSION

As polyadenylation takes place just before the end of transcription, the ability to accurately predict a polyadenylation site would be useful in predicting the ends of transcripts and also terminal exons. With both my human and arabidopsis prediction models having achieved reasonable accuracy, they would certainly be useful for better gene annotations.

FUTURE DIRECTIONS

There are two main challenges while doing this UROP project.

- Finding the "correct" set of features to generate is the key to building high accuracy prediction models. However, it is a daunting task due to it being NP-hard.
- While nice software packages exist for general machine learning (e.g., WEKA), these are frequently clumsy to use for sequence analysis and prediction.

In a bid to overcome these challenges, I will focus my Honors Year Project efforts on them.

REFERENCES

[1] Huiqing Liu, Hao Han, Jinyan Li and Limsoon Wong, (2003a). An In-Silico Method for Prediction of Polyadenylation Signals in Human Sequences. Genome Informatics, Vol. 14, 2003, pp. 84-93.

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

[3] 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.