A Generic System For Genomic Feature Recognition

Student: Koh Chuan Hock and Supervisor: Professor Wong Limsoon

National University of Singapore

BACKGROUND

Biologists would usually depend on the prediction from computer systems to determine possible locations of functional sites or possible localization of proteins. From there, they would then design experiments accordingly to validate the predictions. The computational step is important because it would greatly reduce the number of experiments that needs to be carried out subsequently.

Several high quality prediction systems for functional sites have been developed using a general approach. The approach consists of the following sequential steps: 1) Feature Generation, 2) Feature Selection, 3) Feature Integration and 4) Cascade Classifier. However, building such a system is time-consuming and requires specialized skills, and when such a system is built, it is very specific in that it carries out prediction only for a particular functional site of a particular organism.

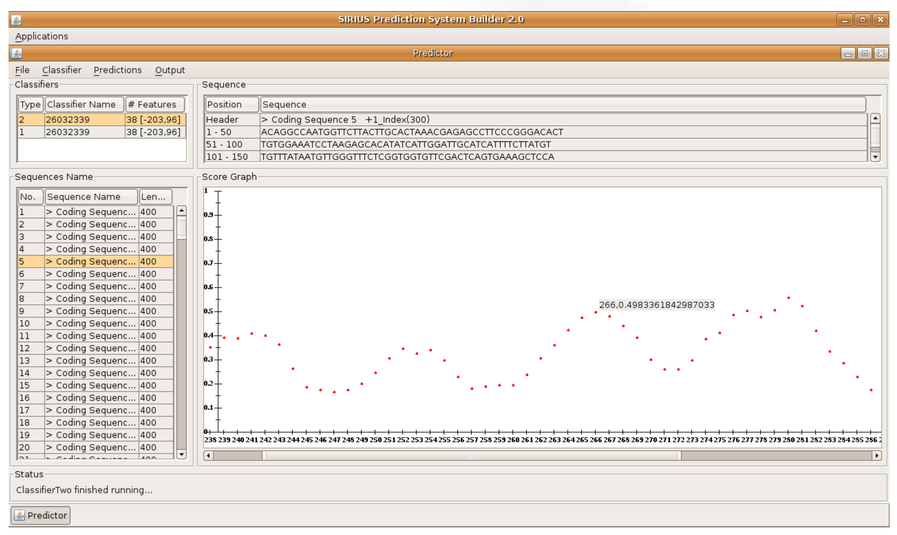

Figure 1. Sirius PSB – Predictor

In an attempt to enable prediction systems to be built, saved and used with ease and speed using the mentioned approach, I developed a software tool named Sirius Prediction System Builder (Sirius PSB). Sirius PSB exudes user-friendliness in that it is equipped with a nice Graphical User Interface that will allow anyone with just some basic knowledge in data mining to be able to build a prediction system without any programming involved. To demonstrate, I have built two prediction models to show the capabilities of Sirius PSB and both models have managed to achieve results comparable to the current state of art.

PREDICTION SYSTEMS

Protein Subcellular Localization: Given a protein sequence, it would be interesting to know the subcellular localization of the protein as it can allow us to better understand its function. Many prediction models have been constructed previously to predict a protein subcellular localization based on its sequence. In particular, TargetP is one such model. It has achieved a high sensitivity (>85%) and is still often used by biologists till today. Hence I will compare my protein localization model against TargetP.

Arabidopsis Polyadenylation Site: Polyadenylation is a post-transcriptional process. This process has been shown to be an essential processing event and an integral part of gene expression. Having the ability to accurately predict them allows us to define gene boundaries, predict the number of genes as well as better understand the process. Currently, the best prediction model for recognition of polyadenylation site for Arabiopsis sequences is designed by Koh et al (2007). Therefore, I will compare my Arabidopsis polyadenylation site model against Koh et al (2007) model.

RESULTS

Protein Subcellular Localization

Table 1. Prediction Performance based on 5-fold cross-validation of TargetP, PL model and Upgraded PL model

			TargetP			PL model			Upgraded PL model		
Set	**Category**	**Size**	**TP**	**FN**	**SN**	**TP**	**FN**	**SN**	**TP**	**FN**	**SN**
Plant	**cTP**	141	120	21	0.851	127	14	0.901			
	mTP	368	300	68	0.815	297	71	0.807			
	SP	269	245	15	0.911	253	16	0.941			
	other	162	137	25	0.846	142	20	0.877			
Plant Sensitivity			0.856			**0.882**					
Non-plant	**mTP**	371	330	41	0.889	344	27	0.927			
	SP	715	683	32	0.955	204	511	0.285	466	79	0.855
	other	1652	1451	201	0.878	1552	100	0.939			
Non-plant Sensitivity			**0.907**			0.717			**0.907**		
Overall Sensitivity			0.878			0.811			**0.892**		

Arabidopsis Polyadenylation Site

Table 2. Equal-error-rate of Koh et al (2007) model and APS model

Control Sequences		**Koh et al (2007) model Sensitivity & Specificity**	**APS model Sensitivity & Specificity**
Coding	**SN_0**	0.943	**0.955**
	SN_10	0.965	**0.971**
	SN_30	0.975	**0.978**
5'UTR	**SN_0**	0.849	**0.854**
	SN_10	**0.892**	0.891
	SN_30	**0.915**	0.912
Intronic	**SN_0**	0.711	**0.724**
	SN_10	0.788	**0.791**
	SN_30	0.830	**0.833**

DISCUSSION

Protein Localization Model: From the results, it is clear that using the approach (feature generation, feature selection and feature integration), comparable if not better performance for the prediction of subcellular localization of proteins can be obtained. Furthermore, in cases when using straight forward features like 1,2 and 3-gram are unable to produce decent results. Sirius PSB also provides automated generation of features using genetic algorithm.

Arabidopsis Polyadenylation Site Model: Although both models followed the same approach, they differed in the candidate features that were generated and how they were selected. For Koh et al, the authors spent a lot of time and effort searching and reading literature about Arabidopsis polyadenylation process to decide on the candidate features. While all I have done here is simply run the genetic algorithm (provided by Sirius PSB) to generate the candidate features. With that, APS model is able to obtain slightly better performance over Koh et al (2007) model.

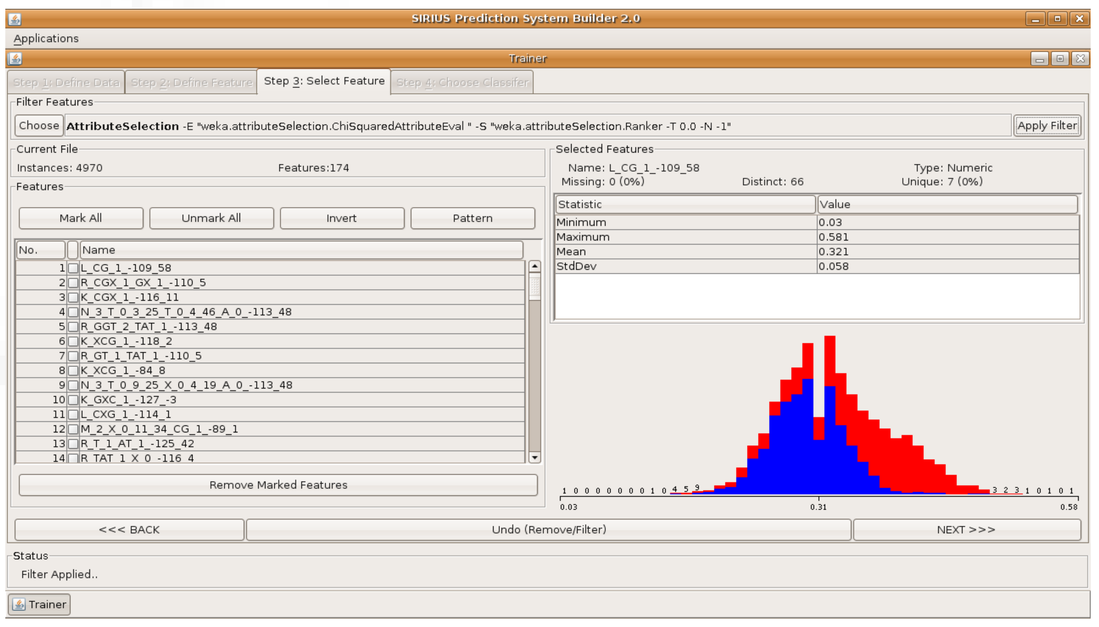

Figure 2. Sirius PSB – Trainer (Feature Selection Step)

CONCLUSION

As shown by the two examples, Sirius PSB has the ability to build high-quality prediction models using the feature generation, feature selection, feature integration and cascade classification approach in a manner that is rapid yet hassle-free. Furthermore, with the genetic algorithm provided in Sirius PSB, users need not even worry about what features to generate.

With Sirius PSB, I am confident that more high-quality prediction models will be produced using the feature generation, feature selection, feature integration and cascade classification methodology.

REFERENCES

[1] Emanuelsson, O., Nielsen, H., Brunak, S. and Heijne, G.V. (2000). Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence. Journal of Molecular Biology, Vol.300, May 2000, pp. 1005-1016.

[2] Koh, C.H. and Wong, L. (2007). Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences. In Proceedings of 18th International Conference on Genome Informatics, (Singapore, December 3-5, 2007), pp. 73-82.

[3] Liu, H. and Wong, L. (2003). Data Mining Tools for Biological Sequences. Journal of Bioinformatics and Computational Biology, Vol.1, No.1, April 2003, pp. 139-167.

[4] Liu, H., Han, H., Li, J. and Wong, L. (2005). DNAFSMiner: A Web-Based Software Toolbox to Recognize Two Types of Functional Sites in DNA Sequences. Bioinformatics, Vol.21, pp. 671-673.

[5] Witten, I.H. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann, San Francisco, 2005.