# Exploring Transcription Patterns in *Arabidopsis thaliana*

**Vishal Bahirwani[1] and Doina Caragea[2]**

## 1 Introduction

Recent work has shown that bidirectional genes (i.e., genes located on opposite strands of DNA, whose transcriptions start sites are not more than 1000bp apart) are often coexpressed and have similar biological functions [5]. Identification of such genes can be useful in the process of constructing gene regulatory networks. Furthermore, analysis of the intergenic regions corresponding to bidirectional genes can help identify regulatory elements, such as transcription factor binding sites.

Wang et al. [6] have identified 2471 bidirectional gene pairs in *Arabidopsis thaliana* and found that the corresponding intergenic regions are rich in regulatory elements that are essential for the initiation of transcription. Identifying such elements is especially important, as simply searching for known transcription factor binding sites [1, 2] in the promoter of a gene can result in many hits that are not necessarily important for transcription initiation. Encouraged by the findings about the presence of essential regulatory elements in the intergenic regions corresponding to the bidirectional genes, in this work, we explore a motif-based machine learning approach to identify intergenic regulatory elements. More precisely, we consider the problem of predicting the transcription pattern for pairs of consecutive genes in *Arabidopsis thaliana* using motifs from AthaMap[3] [1] and PLACE[4] [2], with the goal of identifying the predictive motifs, as those motifs are presumably regulatory motifs that are essential for transcription initiation.

## 2 Direction of transcription for pairs of consecutive genes

We formulate the problem of predicting the direction of transcription for pairs of consecutive genes as a classification problem as follows: given a data set $\mathcal{D} = \{((g_{i,1}, g_{i,2}), c_i)\}_{i=1,\cdots,n}$ of pairs of consecutive genes $g_{i,1}$ and $g_{i,2}$ over the alphabet $\Sigma$ of nucleotides, $|\Sigma| = 4$, $g_{i,1}, g_{i,2} \in \Sigma^*$ along with their class labels $c_i$ that belong to a finite set $C$, the task is to produce a model that is able to predict the class label $c \in C$ for a novel pair of consecutive genes $(g_1, g_2)$. The class label associated with each pair of consecutive genes represents the direction of transcription for the corresponding pair: forward-reverse ($FR$) if the direction of transcription of $g_1$ is forward and of $g_2$ is reverse, reverse-forward ($RF$) if the direction of transcription of $g_1$ is reverse and of $g_2$ is forward, and forward-forward, reverse-reverse ($FFRR$) if the direction of transcription of $g_1$ and $g_2$ is either forward-forward or reverse-reverse.

The *intergenic regions*, i.e. the regions between two consecutive genes on a genome, usually evolve faster than the genes of the genome. However, motifs found in these regions (especially promoter motifs) play significant role in deciding the direction of transcription of adjacent genes and are more conserved. We show how to use these motifs to learn models that can predict the direction of transcription for pairs of consecutive genes and identify the most predictive motifs.

## 3 Motif-based machine learning approaches

Each example in our data set $\mathcal{D}$ is represented as ($100bp\ of\ gene_1$, *intergenic region,* $100bp\ of\ gene_2$).We first collected the motifs from two different sources: $AthaMap$[3] and $PLACE$[4]. $AthaMap$ identifies 117 motifs and $PLACE$ identifies 469 motifs for the *Arabidopsis thaliana* genome. We then encoded each example using the *bag of motifs* representation [3], that is, for each $AthaMap$ and $PLACE$ we constructed a separate data set where each example is a vector of 117 positions and 469 positions, respectively. Each position in the vectors represents the number of times the corresponding motif appears in a given example.

We used various machine learning algorithms (Naïve Bayes Multinomial, Support Vector Machines, Random Forest and Logistic Regression) for learning models for each data set.

[1]Department of Computing and Information Sciences, Kansas State University, USA. E-mail: vishalb@ksu.edu
[2]Department of Computing and Information Sciences, Kansas State University, USA. E-mail: dcaragea@ksu.edu
[3]http://www.athamap.de/
[4]http://www.dna.affrc.go.jp/PLACE/

# 4 Experimental results

We performed our experiments on the *Arabidopsis thaliana* genome that consists of five chromosomes with a total of 30170 gene pairs (extracted from all chromosomes). We used Weka [7] implementations of the algorithms mentioned above to learn classifiers that can distinguish among the three classes considered ($FR$, $RF$ and $FFRR$). Our results show that Random Forests and Support Vector Machines perform the best among all the classifiers that we have explored. Furthermore, class $FR$ is the easiest to learn, followed by $RF$ and finally $FFRR$, for both $AthaMap$ and $PLACE$ motifs (Figure 1).

To identify the most predictive motifs and therefore the most significant regulatory elements, we performed feature selection using the information gain criterion [3]. More precisely, we ordered the lists of $AthaMap$ and $PLACE$ motifs, respectively, using the information gain criterion. Figure 1 shows how the performance of the algorithms (Random Forests and Support Vector Machines) measured through the AUC value, varies with the number of features selected.

Table 1 shows the five most predictive motifs for our classification task from both $AthaMap$ and $PLACE$, respectively. A brief literature review shows that indeed these motifs correspond to regulatory elements that are known to be essential for transcription initiation.

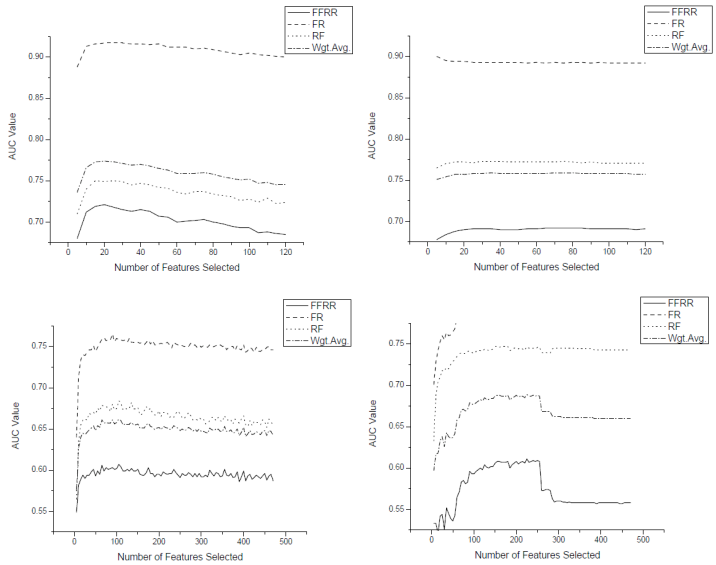


Figure 1: The Area Under the ROC Curve as a function of the number of features selected for both Random Forests (left plots) and Support Vector Machines (right plots) using $AthaMap$ (upper) and $PLACE$ motifs (lower plots), respectively. Using a relatively small number of features (motifs), the classifiers achieve highest performance. As we add more and more features, the performance of classifiers decreases or remains the same.

| **AthaMap Motifs** | CBF | TBP | GT-3B | NTERF2 | DOF2 |
|---|---|---|---|---|---|
| **PLACE Motifs** | GT1CONSENSUS | ARR1AT | POLLEN1LELAT52 | DOFCOREZM | GT1GMSCAM4 |

Table 1: Five most predictive motifs for both $AthaMap$ and $PLACE$.

# References


[1] Galuschka, C., Schindler, M., Bulow, L., and Hehl, R. 2007. *AthaMap web tools for the analysis and identification of co-regulated genes.* Nucleic Acids Res. 35: D857-862.

[2] Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. 1999. Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Research* Vol.27 No.1 pp. 297-300.

[3] McCallum, A. and Nigam, K. 1998 *A Comparison of Event Models for Naive Bayes Text Classification.* AAAI-98 Workshop on "Learning for Text Categorization".

[4] Swarbreck, D. et al. *The Arabidopsis Information Resource (TAIR): gene structure and function annotation.* Nucleic Acids Res 2008, 36:D1009-D1014.

[5] Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P. and Myers, R.M. 2004. An abundance of bidirectional promoters in the human genome. *Genome Research*, 14(1):62-66.

[6] Wang, Q., Wan, L., Li, D., Zhu, L., Qian, M. and Deng, M. 2009. Searching for bidirectional promoters in Arabidopsis thaliana, *BMC Bioinformatics*, 10(1).

[7] Witten, I.H. and Eibe, F. 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.