# User's Guide

Charlotte Darby, Yue Lu, Tony Tong
Bioinformatics Data Integration Practicum 2016

This bioinformatics pipeline is designed to predict transcription termination sites (TTS) in *Zea mays*. First, the user chooses a set of known TTS as the pattern. Then, this pipeline employs motif discovery using MEME (Multiple EM for Motif Elicitation) [Bailey and Elkan, 1994] and the random forest algorithm for machine learning [Breiman, 2001] to characterize the pattern. Once the pattern has been learned, the model can be used to predict new TTS in the maize genome.

## Choose the TTS set

Our method of pattern characterization with MEME and random forests is possible on sets of about 200-600 termination sites. Visit `http://ensembl.gramene.org/biomart/martview/` to get the input sequences for the pipeline [Kinsella et al., 2011]. Choose "Plant genes 50" as the database and "Zea mays genes (AGPv3 (b5))" as the dataset.

The dataset must be filtered so as to obtain 200-600 termination sites. Use a method such as GO terms to choose a set of genes that are more likely to have common regulatory elements near the TTS than a random set of termination sites. In the "Filters" panel, enter the name(s) or accession(s) of the GO term(s) chosen.

We characterize the patterns in the 300bp upstream and 100bp downstream of the end of a TTS, as characterized by the final base pair of a cDNA sequence. In the "Attributes" panel, choose cDNA sequences with 100 downstream flank. Ensure that the search returns the proper number of sequences by using "Count" and then obtain "Results" in a FASTA file.

## Prerequisites

MEME 4.11.1 and Python 2.7 must be installed. The Python packages used are pandas, pickle, Bio, numpy, and sklearn. MEME can be obtained from `http://meme-suite.org/meme-software/4.11.1/meme_4.11.1.tar.gz`.

## Use the pipeline to characterize the TTS set

1. Run `./splitsegments.sh` to prepare the dataset. The first argument is the file path of the FASTA file downloaded from BioMart. The second argument is a short name describing the dataset, e.g. photosynthesis. Type these arguments after the command `./splitsegments.sh`. New directories and files will be created where `./splitsegments.sh` is run; there must be write permissions in this location.

2. After a dataset has been prepared with Step 1, run `./main.sh` to compute features and build a classifier. MEME will be run at this step. The first argument is the short

name of the dataset chosen in Step 1. The second argument is the path to your MEME installation, e.g. `/meme/bin/meme`. Type these arguments after the command `./main.sh`. Files will be created for motifs and to build the random forest classifier; again, there must be write permissions in the directory where `./main.sh` is run.

## Find new examples of the TTS pattern

Based on the input sequences to the previous step, a pattern has been characterized surrounding the termination sites. Now new TTS can be found in the maize genome. Download the repeat-masked version of genome 3.31, found at `ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/zea_mays/dna/Zea_mays.AGPv3.31.dna_rm.genome.fa.gz` [Law et al., 2015]. Alternatively, individual repeat-masked chromosomes can be downloaded from the directory `ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/zea_mays/dna/`. According to Ensembl, in the rm version of the genome/chromosomes, "interspersed repeats and low complexity regions are detected with the RepeatMasker tool and masked by replacing repeats with Ns." Since the maize genome is full of non-coding repeats, this decreases the search space for new TTS to regions of the genome that may contain genes.
3. Run `python finalPrediction.py` to make predictions. The first argument is the short name of the dataset chosen in the first step. The second argument is the file path of the genome or chromosome file downloaded from Ensembl; type these arguments after the command `python finalPrediction.py`. A file `TTS.csv` will be created for the predicted sites so there must be write permissions.

## Sample datasets

Three datasets have been downloaded from BioMart for the GO terms *response to salt stress*, *photosynthesis*, and *response to water deprivation* and are included with the program. These files can be input to Step 1 `./splitsegments.sh` with your choice of short name.

# Bibliography

[Bailey and Elkan, 1994] Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Kinsella et al., 2011] Kinsella, R. J. et al. (2011). Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:2825–2830.

[Law et al., 2015] Law, M. et al. (2015). Automated update, revision, and quality control of the maize genome annotations using maker-p improves the b73 refgen_v3 gene models and identifies new genes. *Plant Physiol.*, 167(1):25–39.