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) and the random forest algorithm for machine learning to characterize the pattern. Then, 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. 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

You must have MEME 4.11.1 and Python 2.7 installed. 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

Move `main.sh` and the sequence file downloaded from BioMart to the directory where you have installed MEME. Run `./main.sh` to start the pipeline. The argument is the file name of the FASTA file downloaded from BioMart; type this after the command `./main.sh`. Files will be created for motifs and build the decision tree classifier; there must be write permissions to this directory.

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`. According to Ensembl, in this version "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. Run `SCRIPT` to start the pipeline. The argument is the file name of the genome file downloaded from Ensembl; type this after the command `./SCRIPT`. Files will be created for motifs and build the decision tree classifier; there must be write permissions to this directory.