CM122/222 – Algorithms in Bioinformatics

Project 2 – Transcription Factor Binding Sequence Prediction

Project Goal:

In this project, the goal is to predict whether a sequence of DNA will be bound by a specific transcription factor in a given condition.

Project Description:

For this project, you will be given a set of actual DNA sequences where a transcription factor is bound based on a Chromatin Immunoprecipitation (ChIP-seq) experiment. In addition, you will be provided a set of randomly selected DNA sequences that do not overlap the bound sequences. Sequences from these two sets will be labeled as such.

You will be given a third set of sequences which contains a mix of additional sequences that are bound by the transcription factor or are unbound. The task will be to predict a subset of the sequences that are likely bound.

For the purpose of this project the identity of the ChIP-seq experiment will be kept anonymous.

Sequences overlapping repetitive elements or gaps in the genome assembly will be excluded in this project.

Project Restrictions:

You are not allowed to use external data not provided as the project inputs.

You are also not allowed to use existing motif discovery libraries or libraries designed to make predictions from DNA sequence.

However, use of standard machine learning libraries is allowed for this project.

Project Inputs:

You will be given three input files. One file will contain a set of DNA sequences bound by the transcription factor as determined by a ChIP-seq experiments (bound.fasta). The sequences will be 200bp in length centered at the center of called peaks. A second file will contain a set of randomly selected unbound sequences (notbound.fasta). A third file will contain a mix of unbound and bound sequences that are not labeled and for which you will make predictions (test.fasta).

The files will be in the “fasta” format used in project 1. As described in project 1, each sequence in a fasta file contains an identifier line starting with a “>” and then the sequence on the following lines.

The sequence names in the bound file with be prefixed with the ‘bound’ prefix, in the unbound file ‘notbound’, and the sequences to be predicted ‘seq’

```
>bound1

AATTTGTGTGCTAAGCAGGGAGAGGCCATCTGAGTTCTCCCATGCATTGA

GTAAACAACACGTCCCGACCTCGTGGCAGGTATTCAGTCAGCTGCTGGTG

AGCAACGTGGCTCCTGTCCCTATGAAGCTTAAAGTCTAAGGTAAGGCAGA

CCCTAGGAATTCGGCGCACAGAAGCCTCTCTGCTGTGATGGAGAAGTAAC

>notbound1

AATTTGTGTGCTAAGCAGGGAGAGGCCATCTGAGTTCTCCCATGCATTGA

GTAAACAACACGTCCCGACCTCGTGGCAGGTATTCAGTCAGCTGCTGGTG

AGCAACGTGGCTCCTGTCCCTATGAAGCTTAAAGTCTAAGGTAAGGCAGA

CCCTAGGAATTCGGCGCACAGAAGCCTCTCTGCTGTGATGGAGAAGTAAC

>seq1

GCACAGAATGCACTATATTGCAGGGGCTGGATAAACACGTGGTGGTTGTC

TTTTTACTAACTTTGATATATTCAATTATATATGGGCCTCATTTCTTCCA

AGCAATATAACAAGTTTTATGGAACAGCCTATCAGTGCTATTTGTCTTAA

ACAAGCCTGAGCCATGTCATCCTAAATTTTATTAACCCCTGTGCATAATG
```

For part a, you will also be given a PWM previously discovered on the bound sequences input file for which you can optionally use in making your predictions. The format of this file is a tab delimited text file. Rows correspond to the nucleotides, A, C, G, and T in order and columns correspond to the different positions of the PWM.

Part b, is based on data for a different transcription factor and you will not be given a PWM corresponding to it.

Project Output: The output for this project is the names of the sequence of your top 2000 predictions of sequences actually bound by the transcription factor based on the ChIP-seq experiment. You will provide them in a text file, one per line. Below is a sample output of the first three lines.

```
seq4
seq5
seq9
 …
```

Evaluation: You will be evaluated based on the number of sequences out of your 2000 predictions that are actually bound by the transcription factor in the experiment.

Full Credit for Undergraduate students: 400 out of 2000 predictions correct for part a and 600 out of 2000 predictions correct on part b.

Full Credit for Graduate students: 400 out of 2000 predictions correct for part a and 800 out of 2000 predictions correct on part b.

Extra credit: Extra credit is possible for exceptional performance based on both datasets.

Note: The thresholds are designed such that full credit is possible for part a by ranking sequences based on the maximum PWM score for any motif instance with the provided motif and for part b by doing the same except for an adequately discovered motif. Other approaches or in combination are possible and may lead to improved predictive performance.

Note: Transcription factors can bind both the forward strand (given) and the reverse complement sequence.