Binding site prediction for MYC

Introduction

This problem is based on local alignment and dynamic programming. Various methods can be used, the one I will be shortly present below, but others such as Hidden Markov Model and training a Neural Network can be more sophisticated and time consuming to implement.

Solution

The algorithm is based on dynamic programming and predicts the most likely binding site of the MYC in the human chromosome 1.

MYC, also called c-myc, is constitutively expressed in cancer cells thus is considered te be a proto-oncogene.

The position weight matrix takes in account all the the reads variants and from it we can calculate the consensus sequence, thus the one most likely to occur.

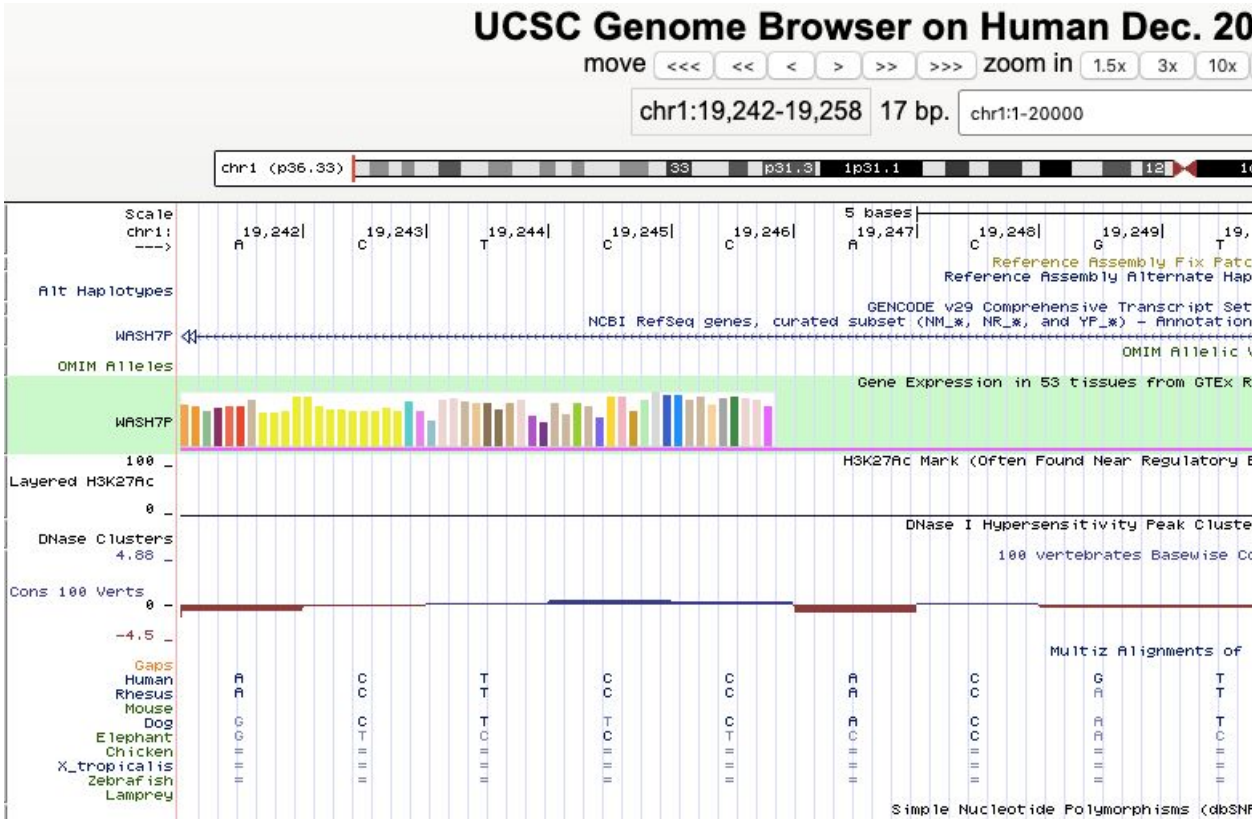
A [	1038	1019	438	44	4600	54	169	59	23	419	961	755	**]**
C [	1345	1287	3746	4840	59	4525	157	162	56	2801	1524	1785	**]**
G [	1649	1757	472	31	170	70	4578	63	4817	1198	1036	1348	**]**
T [	910	879	286	27	113	293	38	4658	46	524	1421	1054	**]**

Once we have our MYC consensus sequence, we can look for it within the human chromosome 1 (only the first 100k bases).

The longest common substring is going to build a matrix with all the matches found. It will output the longest string that is a substring of the reference genome.

GGCCACGTGCCC <--- This is the consensus pattern
  | | | | | | | |
  CCACGTGC <---- Prediction at position <19245> in the reference genome

Looking at the Genome Browser in chromosome 1 at the position 19245 we find our predicted sequence. The gene expression track tells us that those genes (in this area) are likely to be expressed in various tissues, and thus support our prediction of the ‘CCACGTGC’ to be a binding site for MYC .



What was challenging here, was to build a fully reproducible algorithm within the time limit. Due to time and my ‘inner voice telling me you will never get it done in time if you use a more complex approach’ I went on implementing a dynamic programming algorithm.
HMM are known to be good approaches for identifying transcription factor binding sites and predicting the most likely site. I used it once for creating poems using shakespeare style. Next time I will go for this other approach.

The Challenge:

- Develop an algorithm for transcription factor binding event prediction.
- Identify where in the first 100,000 bases of chromosome 1 the MYC transcription factor is most likely to bind.
- Write a report on your results.

The Data:

The humane reference genome chromosome_1 can be downloaded from ensembl:
ftp://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/

A position weight matrix for the MYC binding site can be downloaded from the JASPAR database:
http://jaspar.genereg.net/matrix/MA0147.3/