Assignment 5

Candidate Selection & Putative Cleavage Location

Total Number of reads examined – 42378272

Total Number of candidates selected – 1126

The location of the putative cleavage site for each of the selected candidate can be found in candidates.txt. The putative cleavage location is marked by a dot(.)

Candidates were selected using the following criteria:

1. The reads ends with at least 6 A's
2. Alignment score is less than -6
3. SoftClip is at least 6
4. Number of mismatch is at least 0.

I used a more relaxed criteria and got more than 1500 candidates. Comparing in the genome browser seemed to indicate false positives. So, I decided to use the above conditions. The soft clip value was set to be at least the expected number of A's in the end, so that if read ends in mostly A's, and most of those A's are clipped, it's a promising candidate.

WMM

In the output file for 'wmm 2a' and 'wmm 2b' cases, the following are reported for each WMM0, WMM1, WMM2 :

1. Probability matrix
2. Weight Matrix Model computed from the probability matrix using a uniform background
3. Motif Hit, Log Likelihood Score, Length of the sequence from the beginning of the putative cleavage site.
4. Candidate count with log likelihood score greater than 0.
5. Average distance between putative cleavage site and left end of the best scoring hit with log likelihood greater than 0.
6. Relative entropy of the model.

`computeP2()` - runs the MEME algorithm to calculate the probability matrix. This probability matrix is then used for WMM2 calculation.

`computeWMMFromFreq()` - calculates the weight matrix model using the probability.

`calculate()` – calculates the WMM, log likelihood of all 6-mers and takes the best for each read. If LLR > 0, it is taken as a positive motif hit. The average distance is calculated based on this.

`getRelativeEntropy()` – computes relative entropy of the model.

WMM 2A

Results can be found in wmm_2a.txt. This computation is based on the test data set that was shared in the assignment.

To see the console output of this, please set runToyExample Boolean flag to true in Program.cs of Assignment5 C# Project.

WMM Label	Candidate Count	Candidate count with positive LLR	Average distance	Relative entropy
WMM0	3	0	-	12
WMM1	3	3	39.33	6.914
WMM2	3	3	42	8.196

WMM 2B

Results can be found in wmm_2b.txt

To see the console output of this, please set runToyExample Boolean flag to false in Program.cs of Assignment5 C# Project. If we want to find candidates and then run WMM, please set runFindCandidates Boolean flag to true and ensure that the sam input file path is updated in variable inputSamFile. Otherwise to read candidates from file and then run WMM, set runFindCandidates Boolean flag to false.

WMM Label	Candidate Count	Candidate count with positive LLR	Average distance	Relative entropy
WMM0	1126	212	29.528	12
WMM1	1126	1102	19.260	6.914
WMM2	1126	1090	19.307	8.106