Final Project Report: MotifSeeker

Justin Hii, A16682841

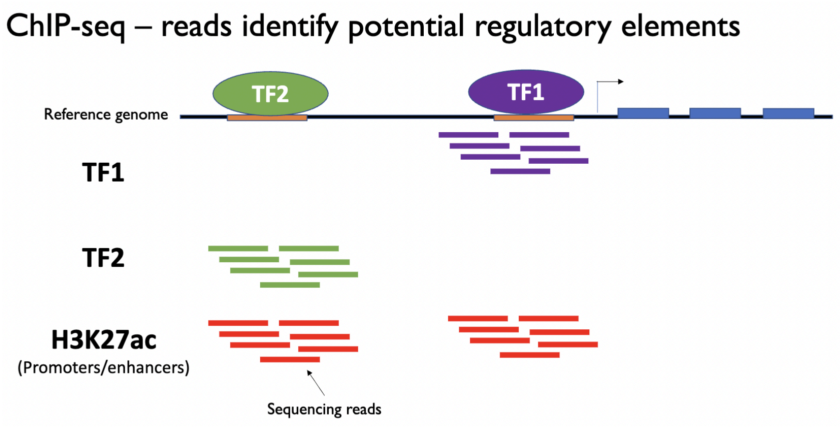
Tusha Karnani, A17339806

Numa Yazadi, A16837717

Name of Tool: MotifSeeker

Introduction:

Motif identification is an important aspect of analyzing gene regulation and expression. It helps in understanding the binding sites of DNA-associated proteins and provides insights into the regulatory elements of the genome. Several layers of computation are performed on peak data obtained after performing ChIP sequencing to identify motifs, which are short, recurring patterns in DNA that are presumed to play significant roles in regulating gene expression. BED files for peaks are obtained after running ChIP sequencing on the gene to identify binding sites for transcription factors and epigenetically modified regions (methylated, acetylated, etc.). Analysis of these files helps recognize motifs that are statistically overrepresented in specific regions of the genome when compared to a background model. Tools like AME and HOMER currently exist to help with the same.



However, current tools are extremely space and memory-intensive and require greater technical expertise to get familiar with. There is a continuous need for improvement in terms of accuracy and adaptability. Motifseeker aims to produce accurate results with a more user-friendly and intuitive approach compared with other related tools.

Methods:

The workflow of the tool utilizes BED files to extract modified sequences (peaks), compares them to corresponding sequences in the genome, computes PWMs, and scans those sequences to find statistically significant motif occurrences.

Our tool uses its ‘ExtractSequencesFromBed’ function to obtain the peaks’ corresponding sequences from the reference genome. It does so by parsing through the fasta files and using the chromosome number and start and end fields in the bed file. The time complexity for this function is O(length of the chromosome) for each line in the bed file, i.e. each sequence. The ‘get_reads’ function parses through and gets reads from the extracted sequences for lengths from 8 to 25 (the minimum and maximum motif lengths of the HOMER motif library, respectively), which two separate functions then generate position frequency matrices (PFMs) and position weight matrices (PWMs) for. The PFMs contain nucleotide frequency information per-nucleotide per-position. The PWMs are then computed to find out the probability of seeing that nucleotide at that position if the distribution were random. It computes a positive, negative or zero value based on whether we are more or less likely to see a particular base at that position.

‘ComputeNucFreqs’ then computes background nucleotide frequencies from the extracted sequences, which ‘RandomSequence’ then uses to generate random sequences and scores them using PWMs to generate background scores. Scoring entails adding up PWM values for the sequence’s nucleotides at the specified positions in order to check the statistical probability and significance of it occurring. The threshold for the desired p-value is obtained based on background scores using ‘GetThreshold’. This function calculates a threshold based on a null distribution, allowing us to determine the significance of the observed scores. The ‘ParseMotifsFile’ function then acts upon the provided file with pre-published and discovered motifs to compare them with the sequences obtained.

We finally find exact matches using the ‘FindExactMatches’ method and perform qualitative analysis using ‘Compute Enrichment’ and ‘MotifEnrichmentAnalysis’ to check whether they are enriched by performing a Fisher exact test to compute the p-value for motif enrichment and comparing the number of sequences (peak and background) that pass the threshold.

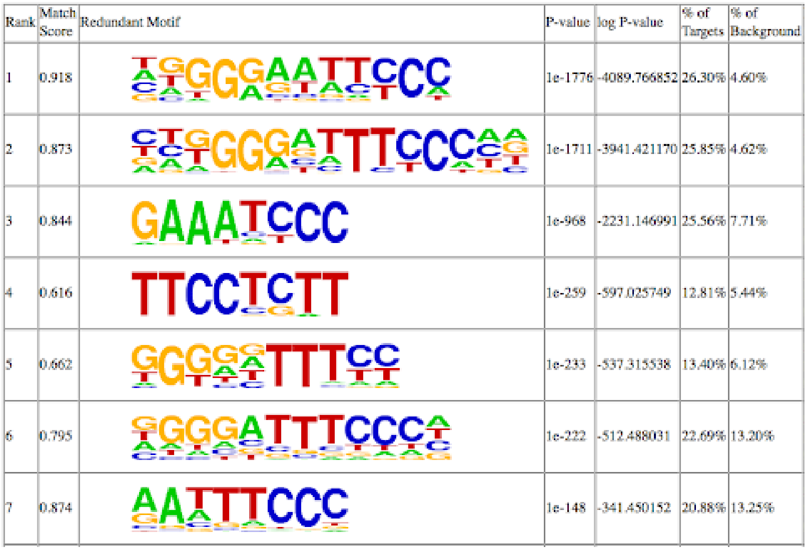
Results:

The output of the file contains a table with the following fields for every significant motif analyzed:

- Serial no.
- Sequence
- Foreground peaks with motif
- Background peaks with motif
- P-value
- Enrichment status (yes/no)

It also looks at motif files with the representations of nucleotides as different letters to signify the appearance of more than one at one position. For example, S for G or C, B or G or C or T, and so on. Our results contain a comprehensive and concise representation of the motifs and their statistical analysis when compared to those found in the genome.

The image below roughly portrays the results that our tool provides, with the fraction of targets and background sequences with the motifs as well.

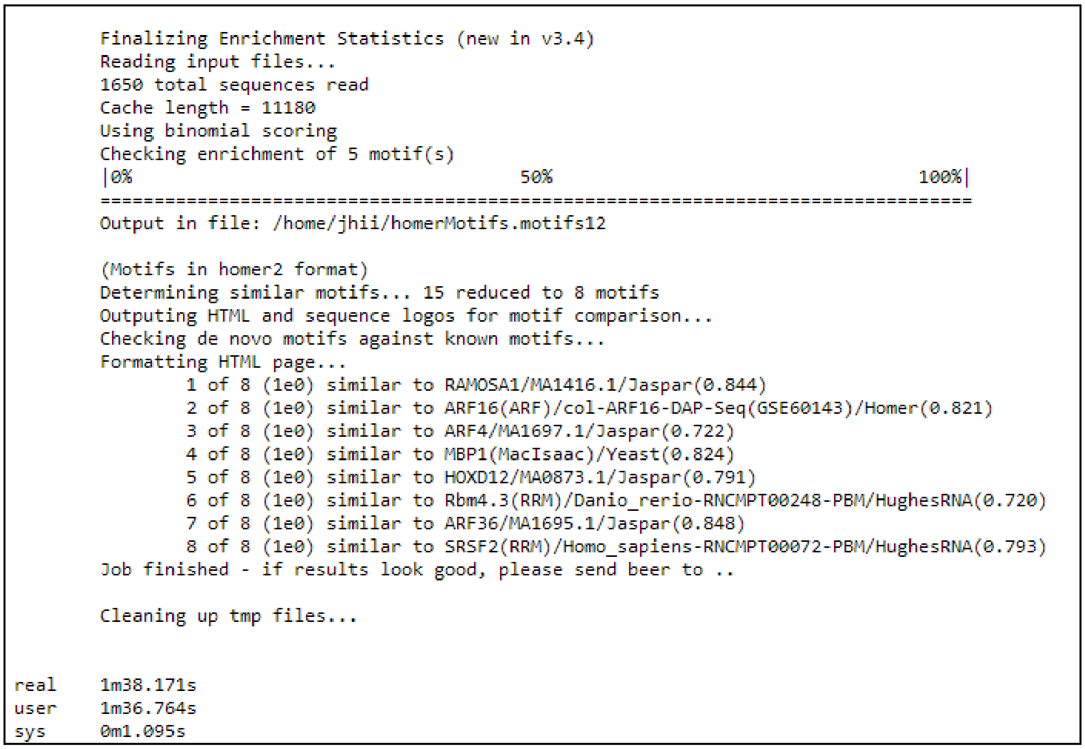


Rank	Match Score	Redundant Motif	P-value	log P-value	% of Targets	% of Background
1	0.918		1e-1776	-4089.766852	26.30%	4.60%
2	0.873		1e-1711	-3941.421170	25.85%	4.62%
3	0.844		1e-968	-2231.146991	25.56%	7.71%
4	0.616		1e-259	-597.025749	12.81%	5.44%
5	0.662		1e-233	-537.315538	13.40%	6.12%
6	0.795		1e-222	-512.488031	22.69%	13.20%
7	0.874		1e-148	-341.450152	20.88%	13.25%

We conducted our benchmarking on a dataset of ChIP-seq experiments, comparing the runtime and accuracy of our tool with HOMER.

To compare the runtime, we measured the time taken by both tools to complete the analysis of the same ChIP-seq dataset. We repeated the tests multiple times to ensure consistency and recorded the average runtimes. To evaluate accuracy, we compared the motifs identified by our tool with those identified by HOMER. We specifically looked at the number of sequences passing the thresholds and the p-values both tools gave us for motif enrichment.

Figure 1 - Here is the output for HOMER (shortened for conciseness):



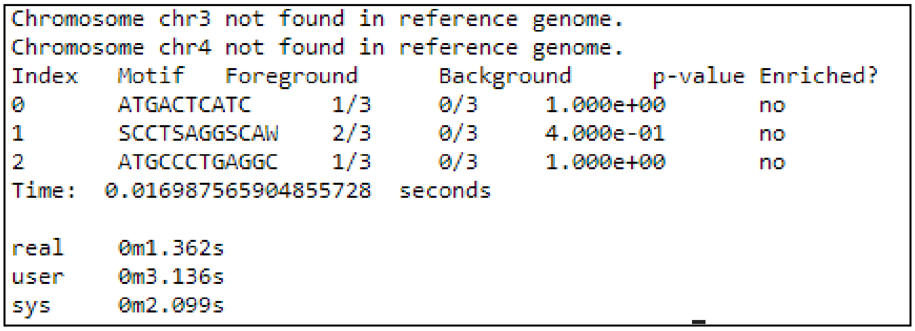
```
        Finalizing Enrichment Statistics (new in v3.4)
        Reading input files...
        1650 total sequences read
        Cache length = 11180
        Using binomial scoring
        Checking enrichment of 5 motif(s)
        |0%                                    50%                                  100%|
        =================================================================================
        Output in file: /home/jhii/homerMotifs.motifs12

        (Motifs in homer2 format)
        Determining similar motifs... 15 reduced to 8 motifs
        Outputing HTML and sequence logos for motif comparison...
        Checking de novo motifs against known motifs...
        Formatting HTML page...
                1 of 8 (1e0) similar to RAMOSA1/MA1416.1/Jaspar(0.844)
                2 of 8 (1e0) similar to ARF16(ARF)/col-ARF16-DAP-Seq(GSE60143)/Homer(0.821)
                3 of 8 (1e0) similar to ARF4/MA1697.1/Jaspar(0.722)
                4 of 8 (1e0) similar to MBP1(MacIsaac)/Yeast(0.824)
                5 of 8 (1e0) similar to HOXD12/MA0873.1/Jaspar(0.791)
                6 of 8 (1e0) similar to Rbm4.3(RRM)/Danio_rerio-RNCMPT00248-PBM/HughesRNA(0.720)
                7 of 8 (1e0) similar to ARF36/MA1695.1/Jaspar(0.848)
                8 of 8 (1e0) similar to SRSF2(RRM)/Homo_sapiens-RNCMPT00072-PBM/HughesRNA(0.793)
        Job finished - if results look good, please send beer to ..

        Cleaning up tmp files...


real    1m38.171s
user    1m36.764s
sys     0m1.095s
```

Figure 2 - Here is the output for MotifSeeker:



```
Chromosome chr3 not found in reference genome.
Chromosome chr4 not found in reference genome.
Index   Motif   Foreground      Background      p-value Enriched?
0       ATGACTCATC      1/3     0/3     1.000e+00       no
1       SCCTSAGGSCAW    2/3     0/3     4.000e-01       no
2       ATGCCCTGAGGC    1/3     0/3     1.000e+00       no
Time:  0.016987565904855728  seconds

real    0m1.362s
user    0m3.136s
sys     0m2.099s
```

We ran both of these tools alongside a “time” argument in order to print their runtimes. Note that the runtime for HOMER was around 1 minute and 38 seconds [Figure 1], while the runtime for MotifSeeker was around 1.362 seconds [Figure 2]. We ran a modified version of ENCFF803UAK.bed (from https://www.encodeproject.org/experiments/ENCSR817LUF/) which was shortened in order to cut down on runtime analysis. The genome we used was GrCH38.fa.

Benchmarking against HOMER gave us important information about our tool's runtime and confidence in its accuracy. `motifseeker` proved to be significantly faster and as accurate as HOMER. Both tools identified similar numbers of sequences passing the thresholds and the p-values for motif enrichment were comparable, with no significant differences observed between the tools.

Discussion:

While our tool performed considerably well on both test and real datasets, we do recognize that it could perform more powerful tests to confirm the validity of its data. It also lacks computational prowess with respect to speed and memory. Here are some ideas that we think could help us improve our tool, given more time.

1. Testing `motifseeker` on larger datasets to ensure it can handle high-throughput ChIP-seq experiments.
2. Incorporating additional functionalities, such as visualization tools for motif analysis results, to enhance user experience.

Our tool has demonstrated superior runtime efficiency and comparable accuracy to HOMER in analyzing ChIP-seq datasets. The promising results motivate us to continue enhancing the tool and expanding its capabilities.

Code availability:

https://github.com/jhii7/MotifSeeker.git

References:

Lee Quian. "Motif Enrichment at Differentially Bound Sites." *ResearchGate*, https://www.researchgate.net/figure/Motif-enrichment-at-differentially-bound-sites-a-Reverse-search-of-known-motifs-at-FLAG_fig3_340289114. Accessed 7 June 2024.

Tran NT, Huang CH. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biol Direct. 2014 Feb 20;9:4. doi: 10.1186/1745-6150-9-4. PMID: 24555784; PMCID: PMC4022013.

Gymrek, Melissa; CSE 185 Lectures Spring 2024