THM | CAMPUS GIESSEN | MNI
TECHNISCHE HOCHSCHULE MITTELHESSEN
Mathematik, Naturwissenschaften und Informatik

Transcription Factor Binding Sites Scan

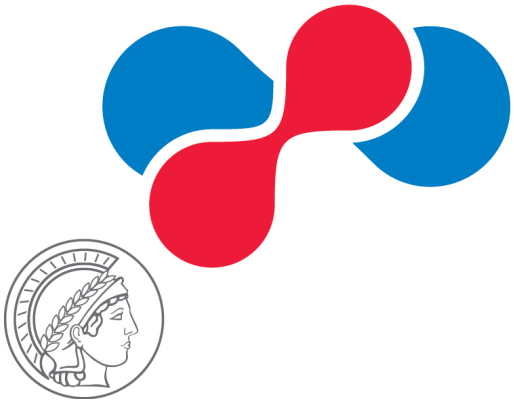
Project Phase

Anastasiia Petrova in THM*

08.05.2018

Supervisors: Dr. Mario Looso, Prof. Dr. Andreas Dominik

Max Planck Institute
for Heart and Lung Research
W.G. Kerckhoff Institute



*FB MNI Mathematik, Naturwissenschaften, Informatik

Transcription factors plays an important role in the gene regulation. To search for the regulatory regions in the genome the Transcription Factor Binding Sites Scan (TFBS scan) was developed. The next few pages contain the brief information about the TFBS scan and the analysis of the TFBS scan.

1 Introduction

By binding to a specific DNA sequence a protein called transcription factor (TF) can control the transcription of genetic information from DNA to messenger RNA. The function of transcription factors is to regulate genes in order to assure the right gene expression in a cell throughout the life cycle. The eukaryotic regulatory regions can be characterized based on a set of discovered transcription factor binding sites (TFBS). The aim of the TFBS scan is to search for these regulatory regions. The produced data is used in the TOBIAS pipeline. TOBIAS is the short name for Transcription factor Occupancy prediction By Investigation of ATAC-seq Signal. The TOBIAS searches for the footprints (regions where the DNA binding protein probably was bound) and the output from the TFBS scan is used to join the footprint and motif information across the genome.

2 Input of the TFBS scan

The TFBS scan requires two input files; they are a file with transcription factor motif and a genome file. The transcription factor motifs could be found in the JASPAR database [1]. There are several formats to download the motifs from the JASPAR database; the TFBS scan works with the .pfm and the .meme formats.

To optimize the scanning user can also pass a file in the .bed format with peaks. The peak is a so called open region or in other words the region of interest in the ATAC-seq. The providing of the .bed file with peaks will make the TFBS scan run way faster and better, as it will look for the needed information only within the regions the user is interested in. Each .bed file has three required columns: the name of the chromosome

(for example, chr1, chrY), the starting and the ending positions. The other columns are optional; they could be the score, the strand information and so on. If the .bed file was provided as input for TFBS scan, the additional information from it will be added to the output of the TFBS scan.

3 Output of the TFBS scan

As an output the TFBS scan produces the file in .bed format, as it is required for the TOBIAS pipeline. If several motifs were given as input of the TFBS scan, user will receive as much output .bed files, as many motifs were given. Each of the output files contains information about only one motif and the file itself has the name of the motif with the motif identification number.

4 The construct of the TFBS scan

There are several steps the TFBS scan takes to produce the output:

1. Check if the input files exist, the input genome file will be also checked if it is in the FASTA format. The other input files will be checked for required formats later when these input files will be needed in different functions. If the file does not exist or has not an appropriative format, the corresponding message will be shown in the terminal and the run of the TFBS scan will stop.

2. If user provides an output directory (see *Customizing the TFBS scan run*), the TFBS scan checks if this directory already exists, and if not, the new directory will be created. if the directory already exists and contains files (for example, from the last run), they will be overwritten. If user has not provided any directory, the new directory called „output“ will be created.

3. If the input file contains several motifs, this input file will be splited into files each containing only one motif.

4. If .bed file with peaks was given as input, this .bed file will be merged with the whole genome file using *bedtools getfasta*. The output of this process is always called „output_merge.fa“ and will be saved in the output directory.

5. The TFBS scan uses multiprocessing to produce the output files. The percentage of the finished work will be printed in the terminal. Here are the steps which multiprocessing will apply to each of the splited motif file:

   a) Call one of the tools: FIMO or MOODS (see *Tools within the TFBS scan*)

   b) Write the output files

   c) Delete temporary files if user wants to

5 Tools within the TFBS scan

User can choose which tool to use within the TFBS scan. Now there are already two possibilities available: FIMO or MOODS. These tools were chosen as they are the most famous and the fastest tools to search for motifs matches within the genome.

5.1 FIMO

FIMO is a part of the MEME Suite [2]. The name FIMO stands for „Find Individual Motif Occurrences“. The program searches a database of sequences for occurrences of known motifs, treating each motif independently. The TFBS scan uses the FIMO to determine all the positions where a transcription factor motif matches in the promoter sequence. FIMO takes motifs in the .meme format as input. From the TFBS scan user can pass additional parameters to FIMO, but the TFBS scan also uses several FIMO options by default, such as --text to print the FIMO output in a simple text file, and --no-qvalue, as while using the --text option the calculation of q-values is not available. Evermore, the q-values are needed neither in the TFBS scan processing nor in the TOBIAS pipeline. Still, if user wants, for example, to do not score the reverse complement DNA strand, it is possible to pass the optional argument --norc from TFBS scan to FIMO (see *Customizing the TFBS scan run*). If FIMO has searched within the merged with peaks genome file, the positions in the FIMO output will be relative to peaks, that is why the TFBS scan sorts the output of FIMO and afterwards counts real positions.

5.2 MOODS

MOODS (Motif Occurrence Detection Suite) is a software package for matching position weight matrices against DNA sequences [3]. MOODS uses advanced matrix matching algorithms implemented in C++. MOODS takes the motif file in the .pfm format as input. The TFBS scan uses MOODS python libraries to produce the output with the same columns as received from FIMO, so that the output of both tools can be treated in the same way later on.

6 Customizing the TFBS scan run

As both available tools within the TFBS scan use motifs in different formats (MOODS uses .pfm and FIMO uses .meme), the TFBS scan offer an automatic converting of these both formats. For example, if a user has provided the motif in .pfm format, but desires to use the FIMO for the computation, the TFBS scan will convert the motifs into the .meme format. To convert the motif from .pfm to .meme TFBS scan uses the small tool provided by the MEME Suite called *jaspar2meme -pfm*. To convert the motif from .meme to .pfm TFBS scan has its own function called convert_meme_to_pfm.

The other problem the user can face is the presence of overlaps. They could cause the enriched regions in the downstream computations within the TOBIAS, that's why it is so important to handle them as soon as we can detect them. The TFBS scan checks the score of overlaps and deletes the ones with worse (lower) score. The regions stayed in the output file after resolving the overlaps have higher score and will be treated as the true matches later in the TOBIAS. Still, the TFBS scan leaves the user to decide if overlaps should be resolved or not, so by default the TFBS scan will not delete any overlaps.

There are several options to customize the TFBS scan:
-o [**directory**]: set the output directory
-b [**file**]: input the .bed file with peaks (see *Input of the TFBS scan*)
--use [**fimo/moods**]: choose the tool to use (by default the TFBS scan uses FIMO)
--clean [**all, nothing, cut_motifs, fimo_output, merge_output, moods_output**]: choose files to delete after the run (by default the TFBS scan will delete all temporary files and will leave only the output .bed files. The user can choose one or more options from the list of deleted files)
--cores [**2**]: choose how many cores you want to use (by default the TFBS scan uses 2 cores)
-p [**0.0001**]: set p-value (by default the TFBS scan uses standard p-value which is 1e-4 or 0.0001)
--resolve_overlaps: delete the overlaps with lower score (by default no overlaps are deleted)
--hide_info: make the run silent, the information will still be written to the log-file (by default all the information is printed to the terminal)
--moods_bg 0.25 0.25 0.25 0.25: set the background to use with moods (by default the TFBS scan uses standard MOODS background which is 0.25 0.25 0.25 0.25)
--fimo „option“: pass the option to FIMO within the quotation marks, for example, set the background file (make the background file for FIMO using *fasta-get-markov* tool from the MEME Suite)

7 Analyzing of the TFBS scan produced data

To check if the TFBS scan works well, several runs of the TFBS scan with different parameters have been done. 54 motifs from the Buenrostros analysis from the cell line GM12878 [4] and the human reference genome hg19 [5] were taken as input. Also the .bed file with peaks from the Buenrostros analysis was taken to make the computation faster and preciser.

To run the FIMO with background the background file was created from the genome file using *fasta-get-markov* from the MEME Suite. Resulted file represents the weights for A, G, C and T within the genome. Background for moods has been set up corresponding to this FIMO background file directly within the TFBS scan as an optional parameter.

All runs represented in this analysis are listed below:

FIMO with overlaps p-value 1e-4	MOODS with overlaps p-value 1e-4
FIMO without overlaps p-value 1e-2	MOODS without overlaps p-value 1e-2
FIMO without overlaps p-value 1e-3	MOODS without overlaps p-value 1e-3
without overlaps p-value 1e-4	MOODS without overlaps p-value 1e-4
FIMO without overlaps p-value 1e-5	MOODS without overlaps p-value 1e-5
FIMO without overlaps p-value 1e-6	MOODS without overlaps p-value 1e-6
FIMO with overlaps, with bg, p-value 1e-4	MOODS without overlaps, with bg, p-value 1e-2
FIMO without overlaps, with bg, p-value 1e-2	MOODS without overlaps, with bg, p-value 1e-3
FIMO without overlaps, with bg, p-value 1e-3	MOODS without overlaps, with bg, p-value 1e-4
FIMO without overlaps, with bg, p-value 1e-4	MOODS without overlaps, with bg, p-value 1e-5
FIMO without overlaps, with bg, p-value 1e-5	MOODS without overlaps, with bg, p-value 1e-6
FIMO without overlaps, with bg, p-value 1e-6	

The results obtained after these runs were compared with the ChIP-seq data [6]. This ChIP-seq data corresponds to the ATAC-seq data from the Buenrostros analysis [4]. The information received from the ChIP-seq shows, where proteins do really bind, while the TFBS scan looks for transcription factors, that can theoretically bind in these regions.

The number of true positive (regions found with the TFBS scan as well as ChIP-seq), false negative (regions found with the CHIP-seq, but not found using the TFBS scan) and false positive (regions found only with TFBS scan, but not with the CHIP-seq) were found.

To measure the accuracy of the TFBS scan several values were chosen [7]:

1. The Sensitivity (the proportion of positives that are correctly identified as such):

$$sensitivity = \frac{TP}{TP + FN}$$ where TP is true positive and FN is false negative

2. The Precision (the positive predictive value):

$$precision = \frac{TP}{TP + FP}$$ where TP is true positive and FP is false positive

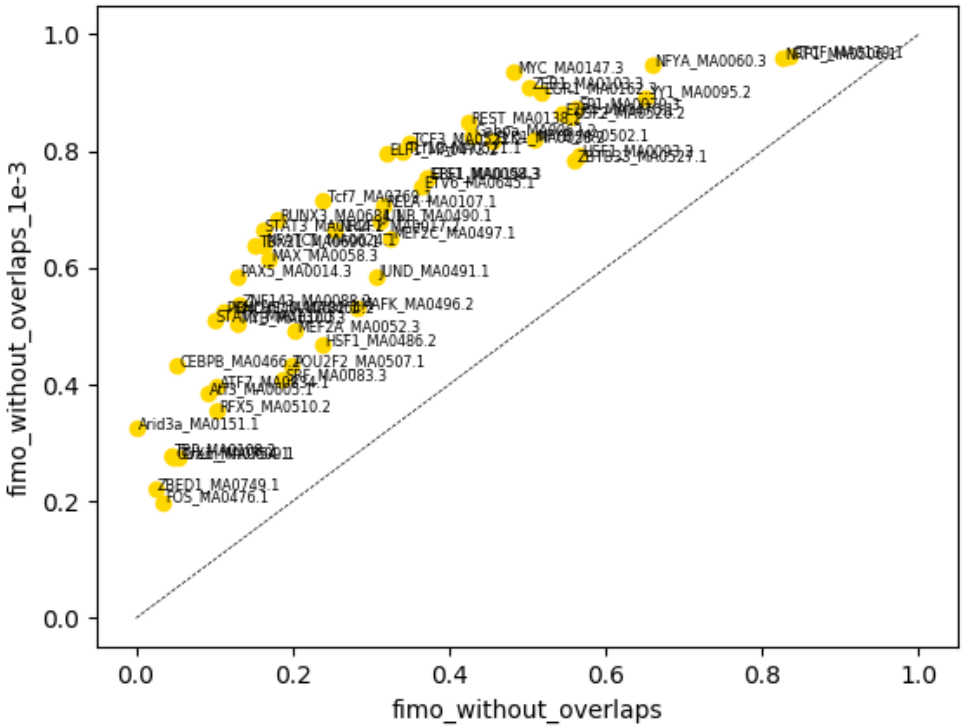
3. The F1-score (the harmonic mean of precision and recall):

$$f1score = \frac{2TP}{2TP + FP + FN}$$ where TP is true positive, FP is false positive and FN is false negative

Also means for all the motifs in each condition was counted to find which condition is the best in overall.

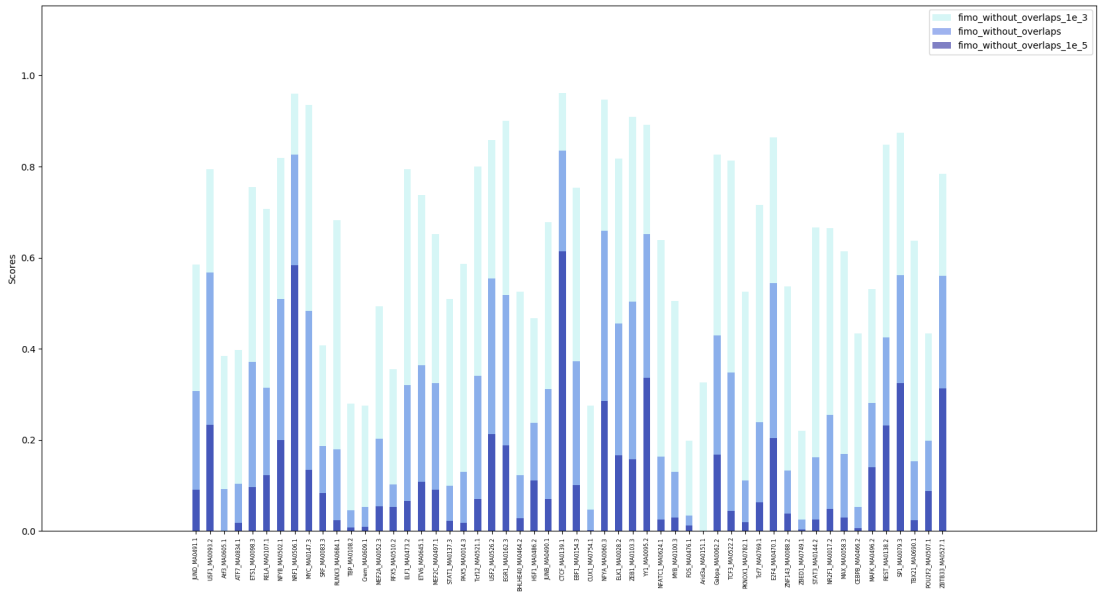
8 Discussion

The picture 1 shows the comparison of sensitivity of FIMO without overlaps, with the standard p-value 1e-4 and FIMO without overlaps, but with the p-value 1e-3. As it is clear to see, for all the motifs the sensitivity is higher with the relaxed p-value. The reason for higher sensitivity is the higher number of found matches. So if it is important for a user to find more possible matches, than it will make sense to set the p-value more relaxed.



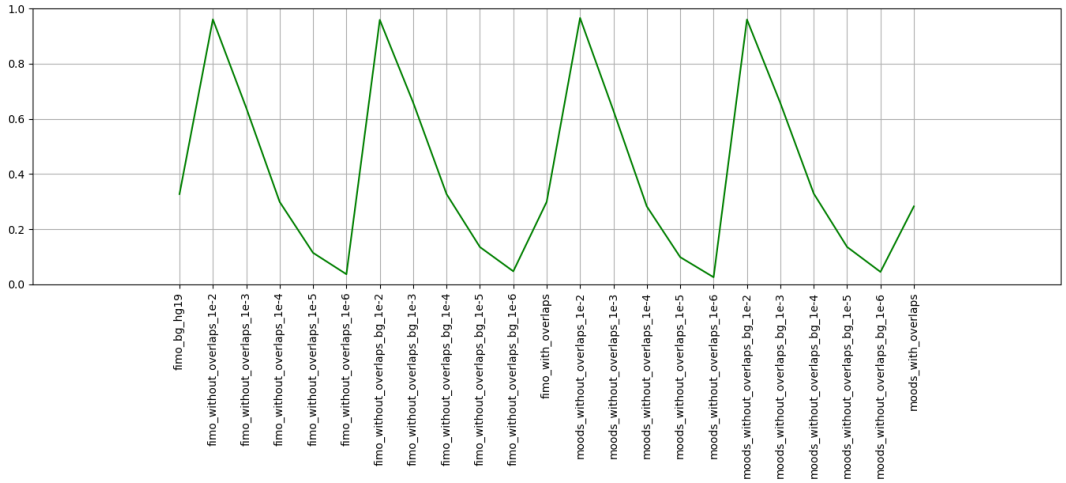
Picture 1: The sensitivity counted in data resulted from runs of the TFBS scan with parameters: FIMO without overlaps, with standard p-value 1e-4 and FIMO without overlaps, with p-value 1e-3

The picture 2 shows the comparison of sensitivity between three FIMO runs. As expected, the sensitivity is higher with the relaxed p-value. The comparison of sensitivity with different p-values of MOODS runs has delivered the same results. For some motifs (for example, Arid3a) it is possible to find any matches only if setting the p-value relaxed. The other motifs (like CTCF) performs well independent of the p-value and the tool used. The conclusion is: the longer the motif is, the better will it perform regarding sensitivity because it is simpler to find this motif. The length of Arid3a is only 6, while the length of CTCF is 19. That is why it is important to know what the length of the motif of interest is, to adapt the p-value for the search of this motif.



Picture 2: The comparison of sensitivity for FIMO runs without overlaps with different p-values

As mentioned already the means for all of motifs in each condition were counted as well. The picture 3 shows the comparison of sensitivity for all runs that were made.

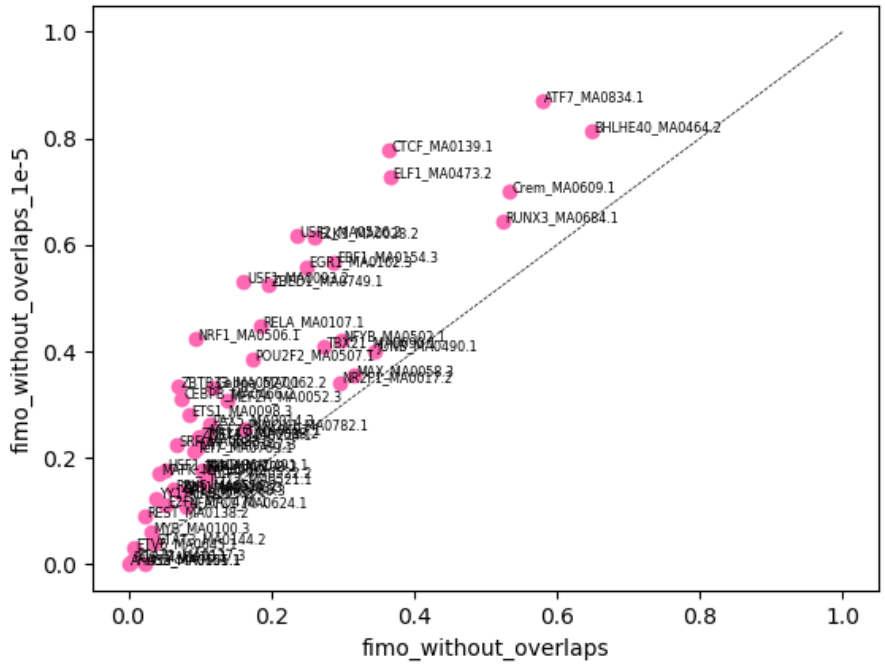


Picture 3: The comparison of sensitivity for all runs

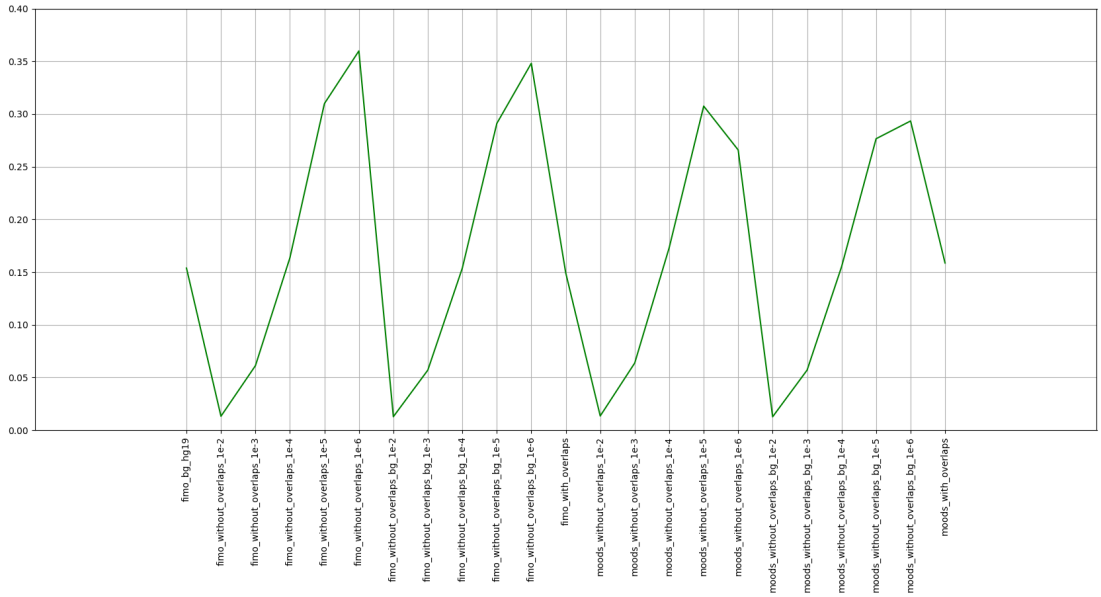
As expected the higher sensitivity can be reached while using the lower p-value just because the TFBS scan will find the higher number of matches.

The picture 4 shows the comparison of precision of runs with different conditions: FIMO without overlaps, with standard p-value (1e-4) and FIMO without overlaps, with

p-value 1e-5. For most of the motifs the strict p-value produces better results regarding precision. The same results could be seen on the picture 5. The lower the p-value is, the higher the precision is.

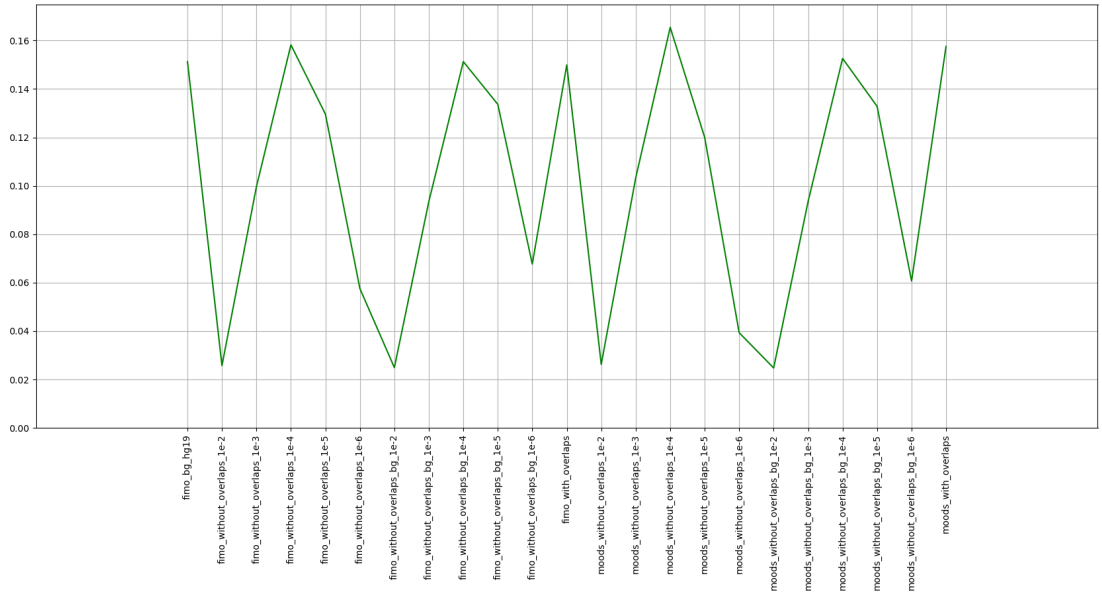


Picture 4: The precision counted in data resulted from runs of the TFBS scan with parameters: FIMO without overlaps, with standard p-value (1e-4) and FIMO without overlaps, with p-value 1e-5



Picture 5: The comparison of precision for all runs

The best way to find the accuracy of the TFBS scan is to count the F1-score. This score is the harmonic mean of precision and recall and it reflects exactly the information we want to know. The picture 6 shows the comparison of F1-score for all runs.

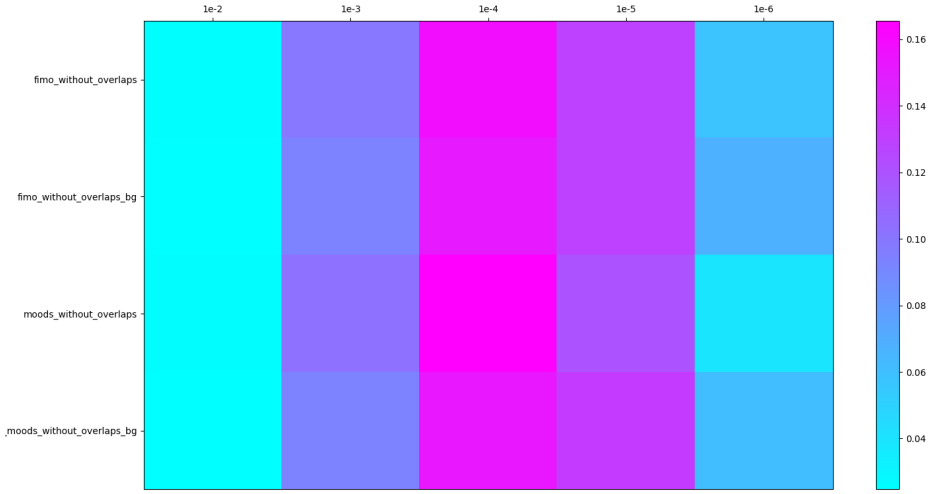


Picture 6: The comparison of F1-score for all runs

The first thing to mention is, that the resolving of overlaps causes better results both for FIMO and MOODS. That is why we excluded testing the data with overlaps and switched to resolving overlaps in all other cases. On the picture 6 only the fimo_bg_hg19, fimo_with_overlaps and moods_with_overlaps represents the runs where the resolving of overlaps was not activated.

It is interesting to see, that the best results were achieved with the standard p-value 1e-4 and while using MOODS, the FIMO with standard p-value has also performed well. The providing of a background will cause better results for stricter p-values in both FIMO and MOODS. In overall FIMO as well as MOODS delivered similar results, which is the reason to say that the quality of result depends much more on parameters set such as resolving the overlaps, setting the background and p-value, than on the tool used.

On the picture 7 the comparison of F1-score is shown in an other way, sorted on the p-value used for the computation. As already mentioned early, the better results are provided with the standard p-value 1e-4. The square corresponding the moods_without_overlaps and standard p-value is the most pink one, what means that the F1-score for these conditions is the the best.



Picture 7: The comparison of F1-score regarding the p-value

It is important for user to know that the analysis shown in comparison with ChIP-seq data depends a lot on the quality of the ChIP-seq data itself. So there could be a probability, that the false positive matches (if TFBS scan has found them, but the ChIP-seq – not) are actually the true positive matches. Still there is no better possibility to test the accuracy of TFBS scan and we should accept the comparison with ChIP-seq data, even if it is not the perfect one.

These are once more the most important observations after the analysis of the TFBS scan:

1. The longer the motif is, the better will it perform because it is simpler to find it
2. The stricter p-value was set, the lower the sensitivity is (as there are less matches found) and the higher the precision is (as these are the better matches)
3. The background will provide better results for stricter p-values
4. The resolving of overlaps will make the F1-score rise

9 Conclusion

The question was to find out which conditions are better to use for the given set of input files, and after the analysis we can say, that it makes sense to use the standard p-value, resolving of the overlaps and MOODS, but one can use FIMO as well. These pieces of advice are though meaningful for the whole set, and if using the certain motif on its own, the better conditions could be not the same as mentioned.

The one more step we are thinking to make is to calculate the p-value for the MOODS output as well. The FIMO produces p-value, but the developer of MOODS do not provide the computation of p-values. If there will be the need for p-value within TOBIAS pipeline or among other users of TFBS scan, we will add the computation of p-values for the MOODS output and to the output of the TFBS scan as well.

The current code is available on GitHub: https://github.molgen.mpg.de/anastasiia/TFBSscan

Literatur

[1] The JASPAR database http://jaspar.genereg.net, last accessed on 07.05.2018.

[2] The MEME Suite http://meme-suite.org/, last accessed on 07.05.2018.

[3] The MOODS https://github.com/jhkorhonen/MOODS/wiki, last accessed on 07.05.2018.

[4] J. Buenrostro, P. Giresi, L. Zaba et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position*, Nature Methods, Vol. 10, No. 12, Pages 1213-1218, 2013. The data provided online at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47753, last accessed on 07.05.2018.

[5] The human reference genome could be downloaded from many databases, for example from Ensembl: http://ensemblgenomes.org/info/access/ftp, last acessed on 07.05.2018.

[6] The ChIP-seq data: https://www.encodeproject.org/search/?type=Experiment&assay_title=ChIP-seq&biosample_term_name=GM12878&limit=all&target.investigated_as=transcription+factor, last accessed on 07.05.2018.

[7] The confusion matrix: https://en.wikipedia.org/wiki/Confusion_matrix, last accessed on 07.05.2018.