# Motif Finding

## CSE463 - Assignment

*Prepared by:*
Lara Khan - 1905062
Majisha Jahan Disha - 1905089
Kazi Istiak Uddin Toriqe -1905104
Arnob shaha Ankon -1905108
Arko sikder -1905109

*Course Teacher:*
Dr. Muhammad Ali Nayeem
Assistant Professor, CSE, BUET

March 14, 2024

# Contents

# 1 Data

## 1.1 Biomarker

For Finding motif in dna sequence we have used some data sets. Here you will find all the data we have used.

# 2 Methods

To find motifs which is statistically significant we used two well known technique

## 2.1 Randomized Motif Search

### 2.1.1 Description

Randomized motif search methods employ statistical techniques to assess the significance of identified motifs. By comparing the observed motif occurrences in the input sequences to a randomized background model, these methods can determine whether the identified motifs are statistically enriched and unlikely to occur by chance alone. Despite the complexity of motif search algorithms, randomized motif search methods are designed to be computationally efficient. They leverage efficient data structures, optimization techniques, and parallelization to handle large-scale datasets and expedite motif discovery processes.

### 2.1.2 Pseudocode

**Algorithm 1** Randomized Motif Search

```
Require: $DNA$: Set of input DNA sequences, $k$: Motif length, $t$: Number of sequences
Ensure: Best motifs found
$motifs \leftarrow$ randomlytMotifs($DNA$, $k$, $t$)
$bestMotifs \leftarrow$ motifs
while True do
    $profile \leftarrow$ Profile($motifs$)
    for all $sequence$ in $DNA$ do
        $motif \leftarrow$ mostProbableKmer($sequence$, $k$, $profile$)
        add $motif$ to $motifs$
    end for
    if $score(motifs) < score(bestMotifs)$ then
        $bestMotifs \leftarrow motifs$
    else
        return $bestMotifs$
    end if
end while
```

## 2.2 Gibbs Sampler

### 2.2.1 Description

The Gibbs sampler motif search is a popular algorithm used in bioinformatics to identify overrepresented sequence motifs within a set of DNA or protein sequences. The Gibbs sampler is a stochastic algorithm that iteratively samples motif occurrences from the input sequences. It starts with an initial set of motifs and iteratively updates them to maximize the likelihood of observing the input sequences given the motifs.The Gibbs sampler uses a probabilistic model to represent motifs, typically in the form of a position weight matrix (PWM) or profile matrix. This model describes the probability of observing each nucleotide or amino acid at each position within the motif. Gibbs sampler motif search is a powerful algorithm for discovering overrepresented motifs in biological sequences. It leverages stochastic sampling techniques and probabilistic models to identify motifs that are statistically significant and biologically relevant

### 2.2.2 Pseudocode

**Algorithm 2** Gibbs Sampler Motif Search

```
Require: $DNA$: Set of input DNA sequences, $k$: Length of motif, $t$: Number of sequences, $N$ : Number of iterations
Ensure: Best motifs found
$motifs \leftarrow$ randomlytMotifs($DNA$, $k$, $t$)
$bestMotifs \leftarrow$ motifs
for $j = 1$ to $N$ do
    $i \leftarrow$ RandomRange(t)
    motifs.popAtIndex(i)
    $profile \leftarrow$ Profile($motifs$)
    $motif \leftarrow$ mostProbableKmer($sequence[i]$, $k$, $profile$)
    motifs.insertAtIndex(i,motif)
    if $score(motifs) < score(bestMotifs)$ then
        $bestMotifs \leftarrow motifs$
    end if
end for
return Best motifs found
```

# 3 Software

We used two Softwares : RSAT and Homer

## 3.1 Commands to run

### 3.1.1 RSAT

RSAT is an online tool.There's some filter format for visualizing data

- Format => Dataset input format
- Matrix length => The value of K
- Sites per Sequence
- Markov-Order
- Background Model

These are the important parameters of RSAT tool.

### 3.1.2 Homer

Homer is a software based on Perl language. There's an details installation guidelines available in here Pre-requisites:

- C++/C compiler,Perl,GNU make utility

We are documenting the process how we did

- $sudo apt update
- $sudo apt install perl
- $sudo apt install build-essential
- $wget http://homer.ucsd.edu/homer/configureHomer.pl
- $perl configureHomer.pl -install
- $export PATH=/path/to/homer/bin:$PATH

Now for testing all things are running correctly give the following command

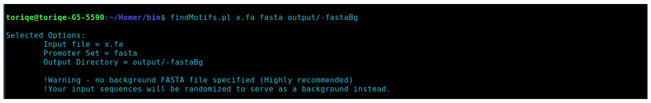
- findMotifs.pl

## 3.2 Scripts to run

### 3.2.1 Homer

We analyzed hm03.txt dataset using an background model to find more statistically significant motifs.The command :

- \$findMotifs.pl x.fa fasta output/-fastaBg



```
toriqe@toriqe-G5-5590:~/Homer/bin$ findMotifs.pl x.fa fasta output/-fastaBg

Selected Options:
        Input file = x.fa
        Promoter Set = fasta
        Output Directory = output/-fastaBg

        !Warning - no background FASTA file specified (Highly recommended)
        !Your input sequences will be randomized to serve as a background instead.
```

We didnot use any background fasta file.For details see here

# 4 Results

## 4.1 Experiment configuration

for each dataset we have run three different programs using 3 different techniques **Varied K from 8 to 24**

### 4.1.1 Program$_1$

- We selected N = 1000 and numberSeed = 50 for progarm1
- We used Minimum hamming distance scoring technique for Randomized Motif Search and entropy technique for Giibs sampler which can be found in git repository from a commit The Entropy Function is as follow :

  $$H(Nucleotide) = -\sum_{i=1}^{n} P(nuocleotide_i) \log_2 P(nucleotide_i)$$

  where $Nucleotide$ is a discrete random variable, $P(nucleotide_i)$ is the probability of the $i$-th outcome, and $n$ is the total number of possible outcomes. And Nucleotide=['A','C','G','T']

### 4.1.2 Progarm$_2$

- We selected N = 1000 and numberSeed = 50 for progarm2
- We used averege Information gain technique for finding motifs on both methods

  $$gain = \frac{maxEntropy * k - \sum_{i=1}^{k} H(Nucleotide)}{k}$$

  where k denotes the motifs length

### 4.1.3 Program3

- We selected N = 100 and numberSeed = 50 for progarm3
- We tested with avg Minimum mismatch technique for both methods

## 4.2 Comparison

### 4.2.1 Progarm$_3$

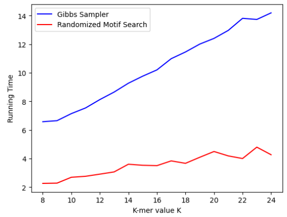


Figure 1: for dataset1

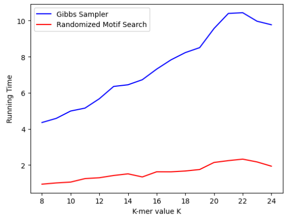


Figure 2: for dataset2

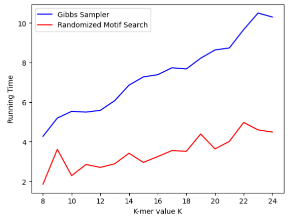


Figure 3: for dataset3

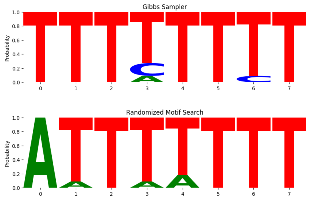


Figure 4: Best Motif Logo of K=8 dataset3

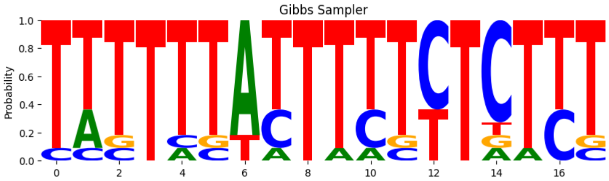


Figure 5: Best Motif Logo of K=16 dataset3

### 4.2.2 Progarm$_2$

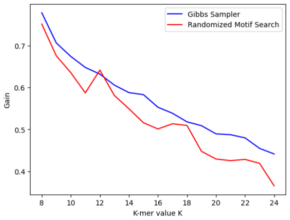


Figure 6: for dataset1

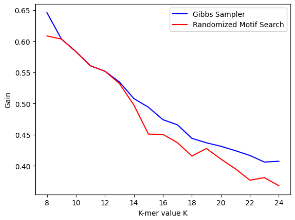


Figure 7: for dataset2

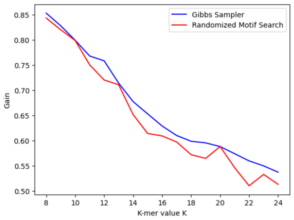


Figure 8: for dataset3

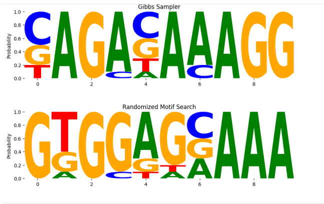
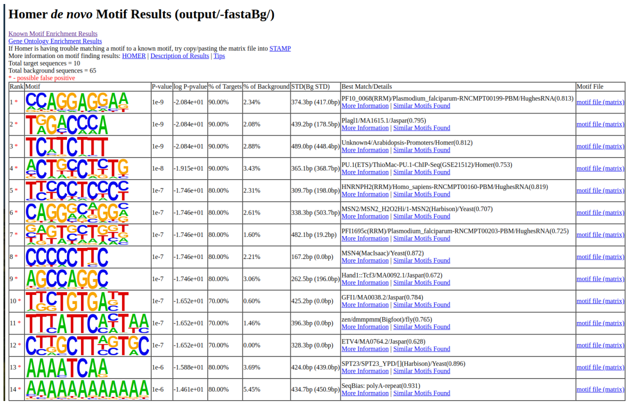


Figure 9: Best Motif Logo of K=10 dataset1

### 4.2.3 Homer and RSAT Output for Dataset1



**Homer *de novo* Motif Results (output/-fastaBg/)**

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 10
Total background sequences = 65
* - possible false positive

| Rank | Motif | P-value | log P-pvalue | % of Targets | % of Background | STD(Bg STD) | Best Match/Details | Motif File |
|---|---|---|---|---|---|---|---|---|
| 1 * | CCAGGAGGAA | 1e-9 | -2.084e+01 | 90.00% | 2.34% | 374.3bp (417.0bp) | PF10_0068(RRM)/Plasmodium_falciparum-RNCMPT00199-PBM/HughesRNA(0.813)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 2 * | TGGACCCA | 1e-9 | -2.084e+01 | 90.00% | 2.08% | 439.2bp (178.5bp) | Plagl1/MA1615.1/Jaspar(0.795)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 3 * | TCTTCTTT | 1e-9 | -2.084e+01 | 90.00% | 2.88% | 489.0bp (448.4bp) | Unknown4/Arabidopsis-Promoters/Homer(0.812)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 4 * | ACTGCCTCTG | 1e-8 | -1.915e+01 | 90.00% | 3.43% | 365.1bp (368.7bp) | PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer(0.753)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 5 * | TTCCCTCCCC | 1e-7 | -1.746e+01 | 80.00% | 2.31% | 309.7bp (198.0bp) | HNRNPH2(RRM)/Homo_sapiens-RNCMPT00160-PBM/HughesRNA(0.819)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 6 * | CAGGGCAGGC | 1e-7 | -1.746e+01 | 80.00% | 2.61% | 338.3bp (503.7bp) | MSN2/MSN2_H2O2Hi/1-MSN2(Harbison)/Yeast(0.707)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 7 * | GAGTGCTGGT | 1e-7 | -1.746e+01 | 80.00% | 1.60% | 482.1bp (19.2bp) | PFI1695c(RRM)/Plasmodium_falciparum-RNCMPT00203-PBM/HughesRNA(0.725)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 8 * | CCCCCTTC | 1e-7 | -1.746e+01 | 80.00% | 2.21% | 167.2bp (0.0bp) | MSN4(MacIsaac)/Yeast(0.872)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 9 * | AGCCAGGC | 1e-7 | -1.746e+01 | 80.00% | 3.06% | 262.5bp (196.0bp) | Hand1::Tcf3/MA0092.1/Jaspar(0.672)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 10 * | TTCTGTGATT | 1e-7 | -1.652e+01 | 70.00% | 0.60% | 425.2bp (0.0bp) | GFI1/MA0038.2/Jaspar(0.784)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 11 * | TTTATTCACTAA | 1e-7 | -1.652e+01 | 70.00% | 1.46% | 396.3bp (0.0bp) | zen/dmmpmm(Bigfoot)/fly(0.765)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 12 * | CTTGCTTAGTGC | 1e-7 | -1.652e+01 | 70.00% | 0.00% | 328.3bp (0.0bp) | ETV4/MA0764.2/Jaspar(0.628)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 13 * | AAAATCAA | 1e-6 | -1.588e+01 | 80.00% | 3.69% | 424.0bp (439.0bp) | SPT23/SPT23_YPD/[](Harbison)/Yeast(0.896)<br>More Information \| Similar Motifs Found | motif file (matrix) |
| 14 * | AAAAAAAAAAAA | 1e-6 | -1.461e+01 | 80.00% | 5.45% | 434.7bp (450.9bp) | SeqBias: polyA-repeat(0.931)<br>More Information \| Similar Motifs Found | motif file (matrix) |

Figure 10: Homer Result for hm03 dataset

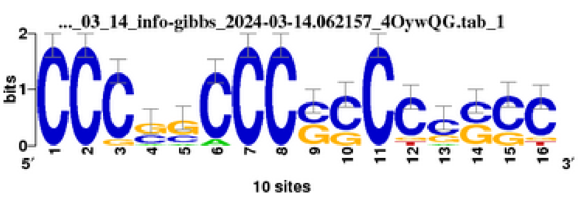


Figure 11: RSAT Result for hm03 dataset, K=16,numSeed=10,iteration=1000

## 5 Conclusion

We see from Figure1, Figure2, Figure3 **Randomized Motif Search** works significantly faster than **Gibbs Sampling**. When we vary the value of K and tends to increase it ,the running time for Gibbs sampler increases. But we can see from Figure6, Figure7 and Figure8 that Gibbs sampler gives much more statistically significant output as the Information-gain is larger and entropy is lower for Gibbs Sampler method.From those plots we also observe that for larger K the gain difference begins to increase than it was before.In the Homer tool testing the results of different K's are ranked.This is because they used some sophisticated scoring technique using mean,deviation,expected values [here] and they used a background fasta file for finding motifs which results statistically more accurate.We tested the RSAT tool output entropy and it was a little bit larger than our implementation result. Homer and RSAT both uses Gibbs Sampler by default.

## 6 References

- RSAT Article
- RSAT GibbsSampler
- Homer Motifs
- Homer Article
- Datasets