Fast and exact motif discovery using the SeqAn library GenMap algorithm

presented by:

Gergana Stanilova

Department of Mathematics and Computer Science

Freie Universität Berlin

Advisor: Prof. Dr. Knut Reinert

Second examiner: Prof. Dr. Alexander Bockmayr

Introduction & Motivation

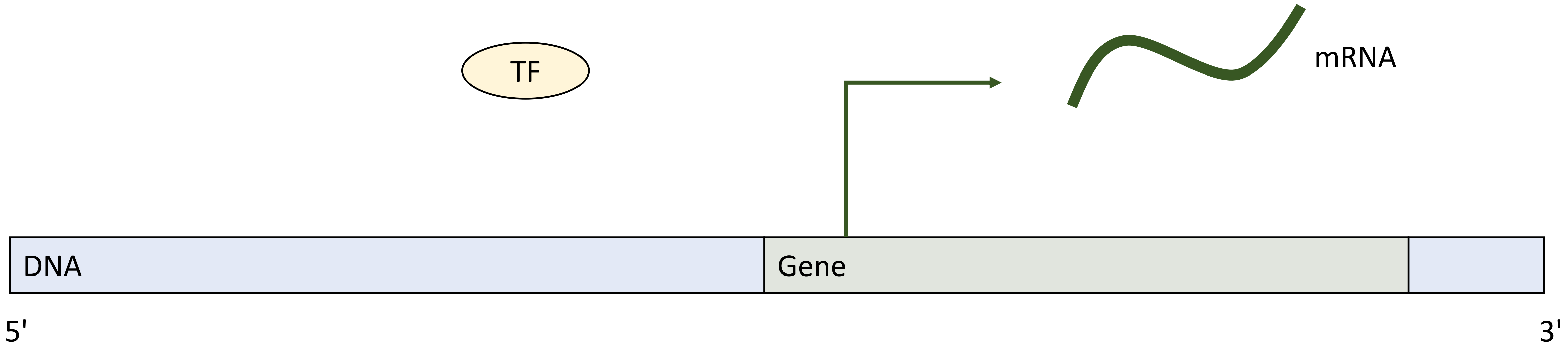
Motif discovery

Identifying recurring patterns within a dataset

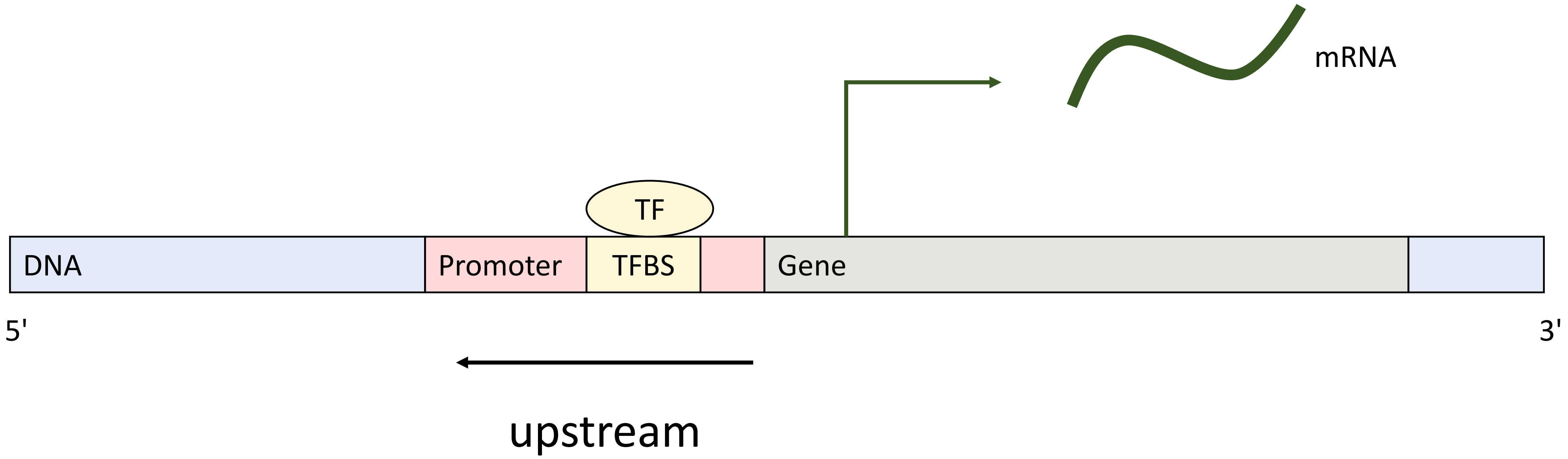
Practical applications of DNA motif discovery

- Understanding gene regulation
- Disease diagnosis and prediction
- Drug target identification

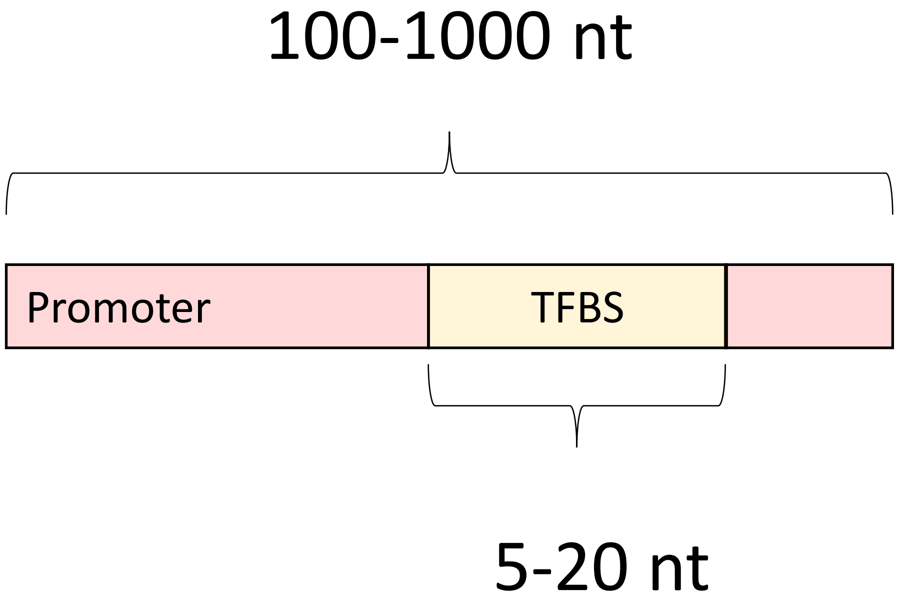
Gene transcription is controlled by a transcription factor (TF).



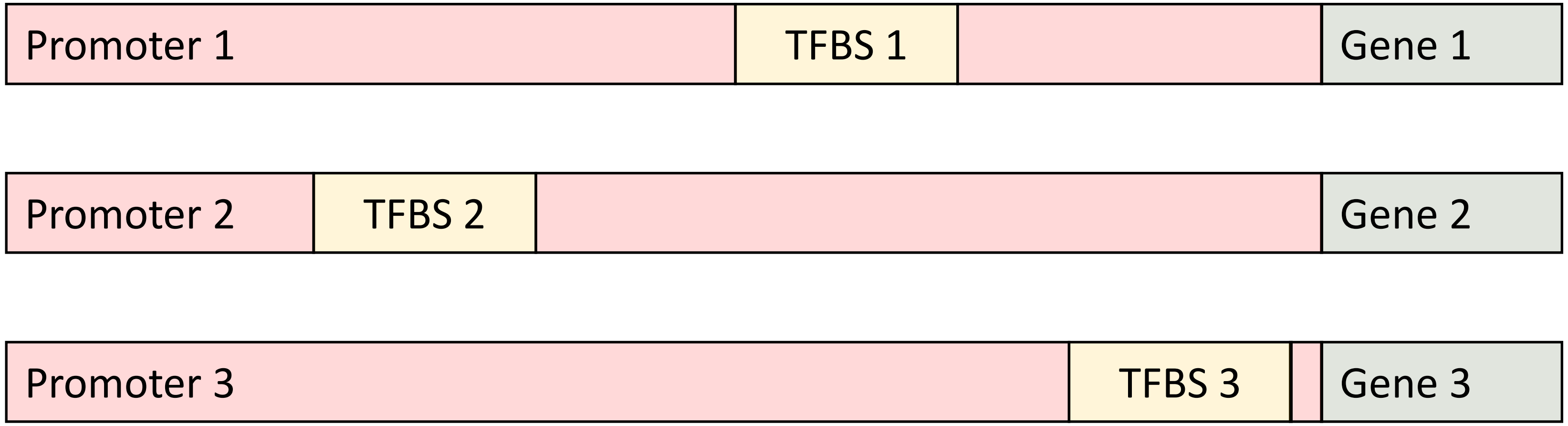
The TF binds to the transcription factor binding site (TFBS) in the promoter region upstream of the gene.



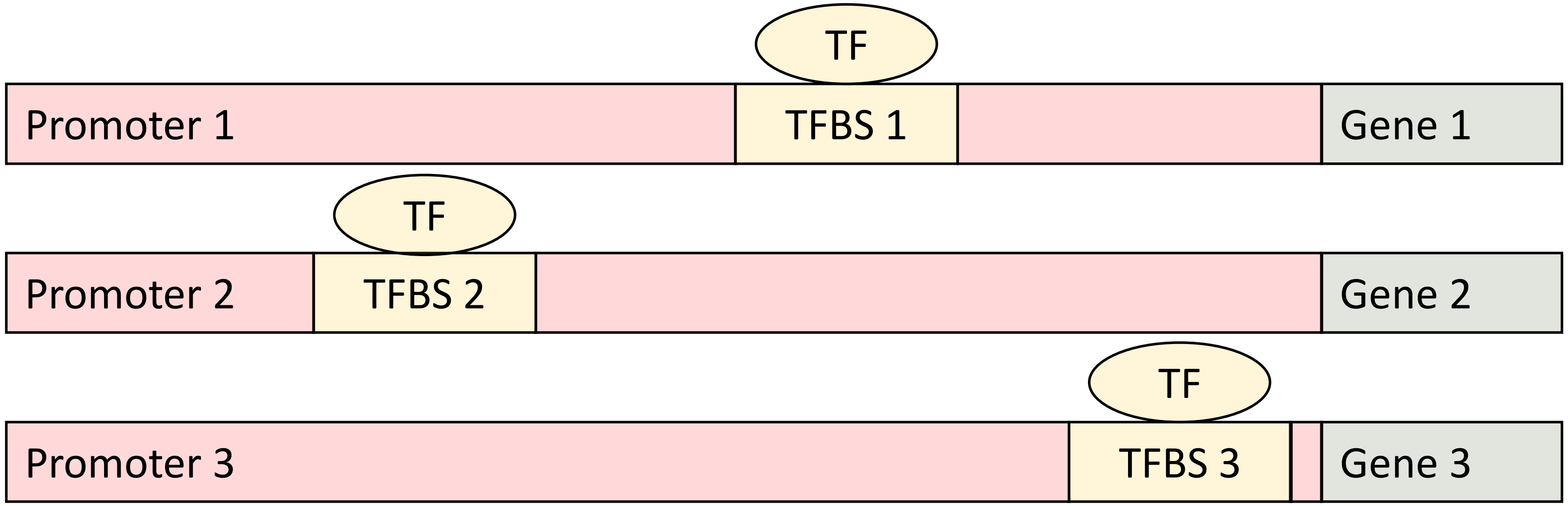
Average length



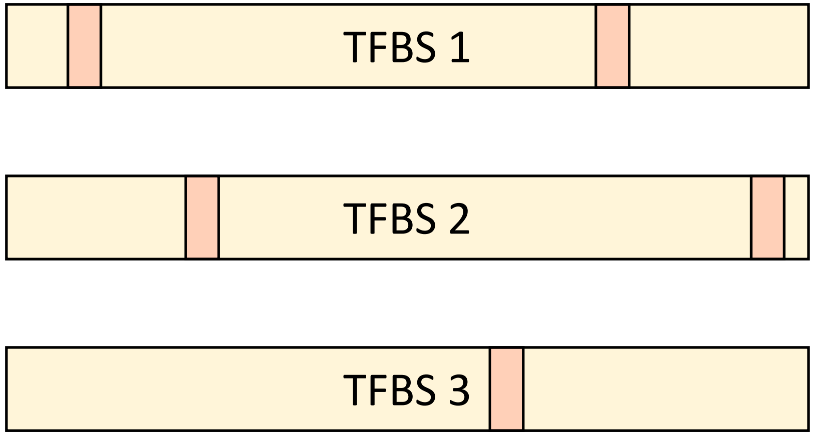
Gene families



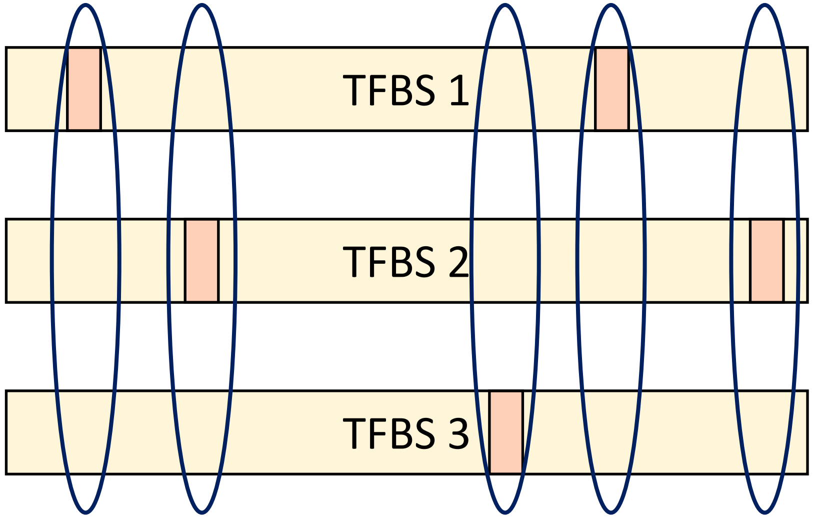
Multiple TBFSs but one TF



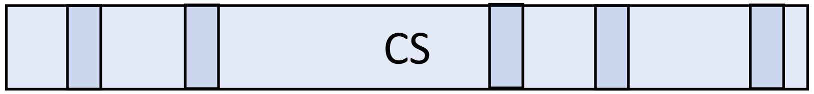
Mutations



Consensus Sequence



Consensus Sequence

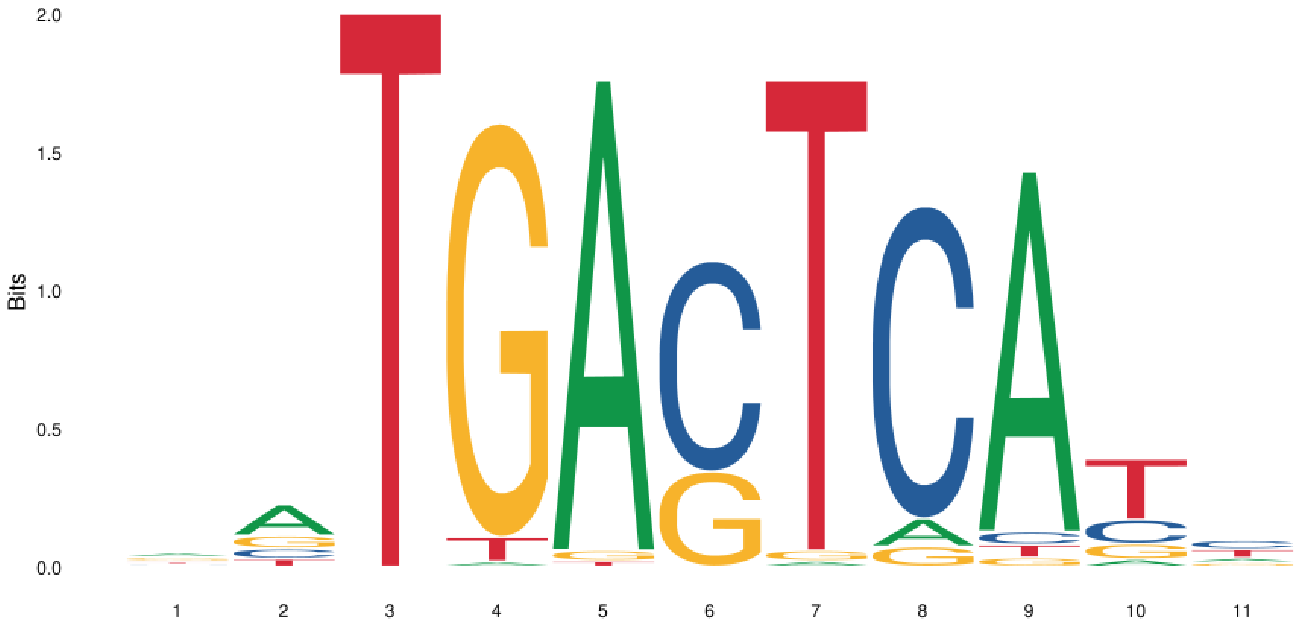


Most common nt at a
certain position

Consensus Sequence of Gcn4

5' TGACTC 3'

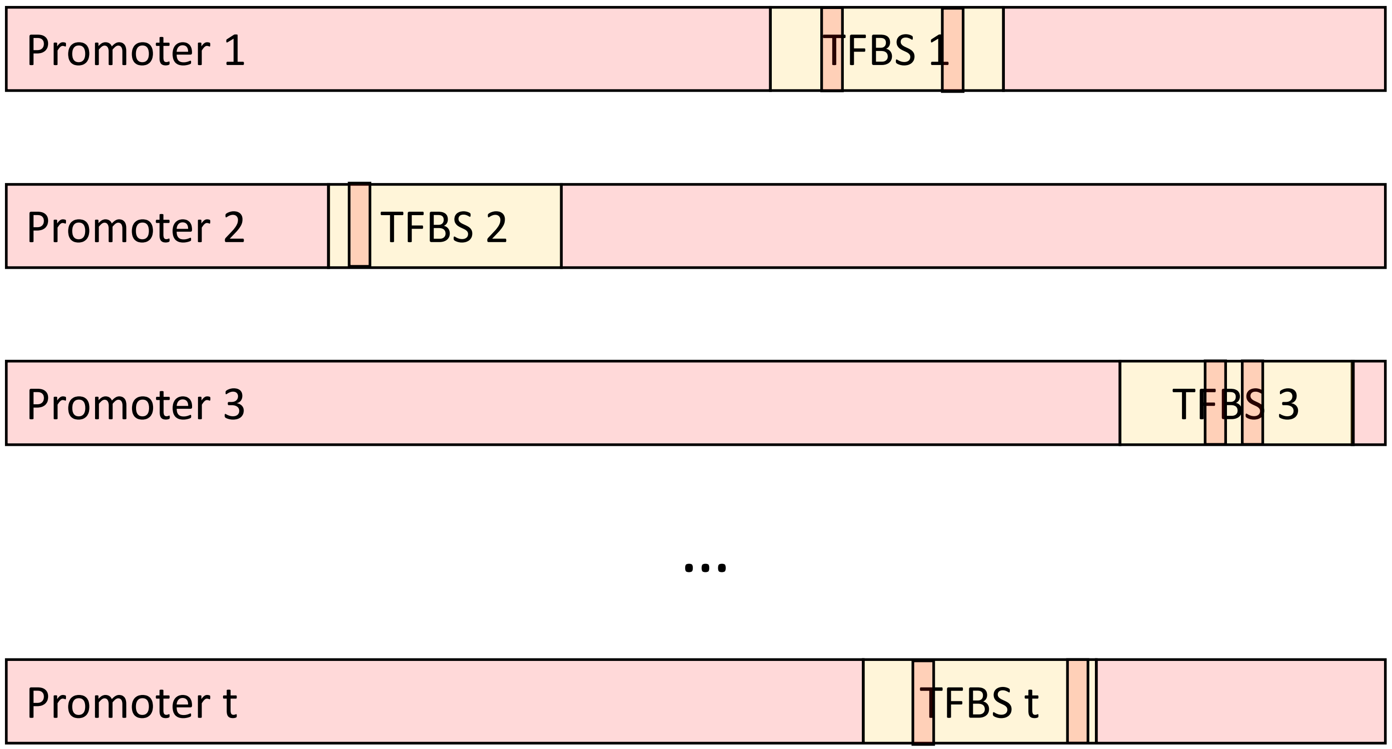
Sequence Logo of Gcn4



Relative sizes of the letters indicate their frequency

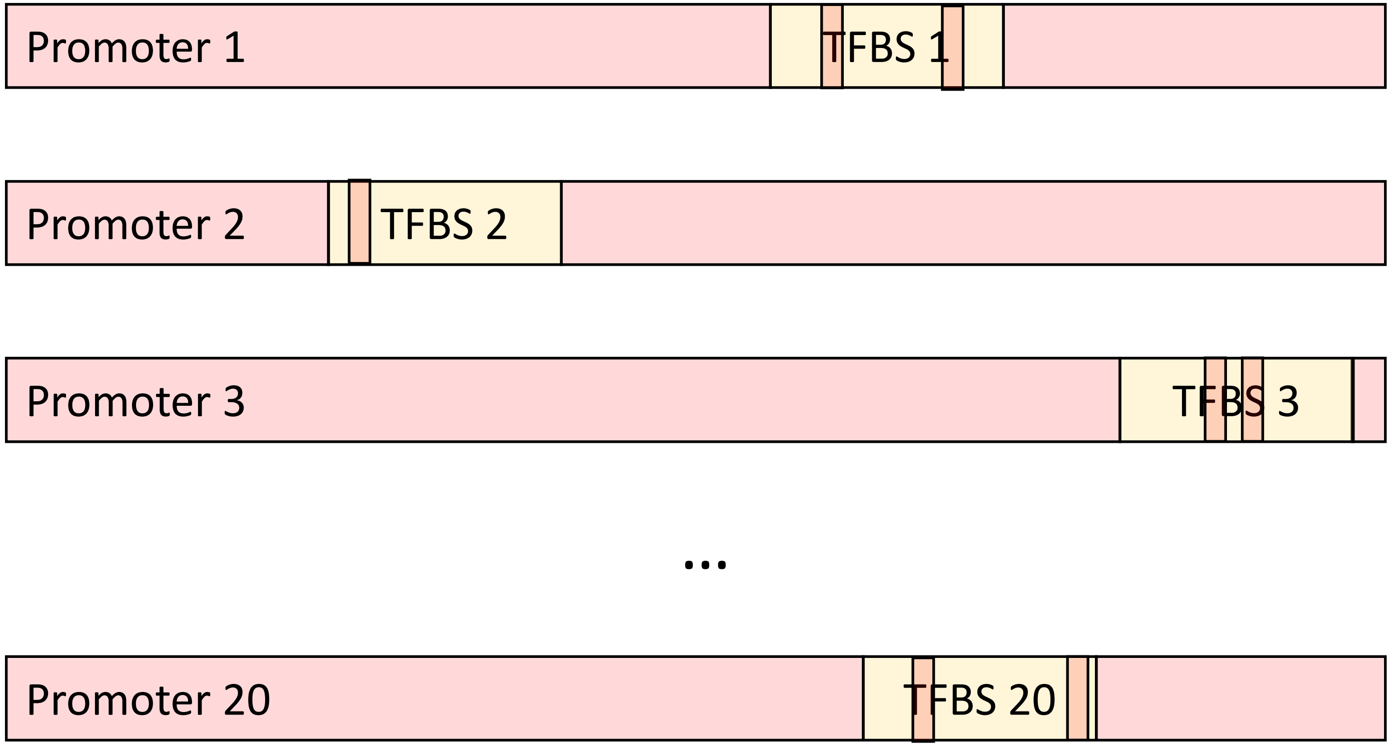
*Image source: https://jaspar.genereg.net/static/logos/all/svg/MA0303.2.svg

The generalized (l,d)-motif problem



In t sequences
of length n
each TFBS is an l-mer
with up to d mutations
with one occurrence
per sequence (OOPS)

The (15,4)-Problem



In **20** sequences
of length **600**
each TFBS is an **15**-mer
with up to **4** mutations

Types of approaches to solving the problem

Exhaustive

- ⊕ suitable for short patterns
- ⊕ no motifs get overlooked
- ⊖ long computation times

Heuristic

- ⊕ fast computation times
- ⊕ practical for long motifs
- ⊖ accuracy depends on initialization

Algorithms used in this thesis

GenMap

Christopher Pockrandt, Mai Alzamel, Costas S. Iliopoulos, and Knut Reinert. Genmap: Fast and exact computation of genome mappability. bioRxiv. 2019. doi: 10.1101/611160

Strategy:

- Performs a linear search
- Does not process every single l-mer separately
- Searches for approximate matches
- Skips redundant l-mers

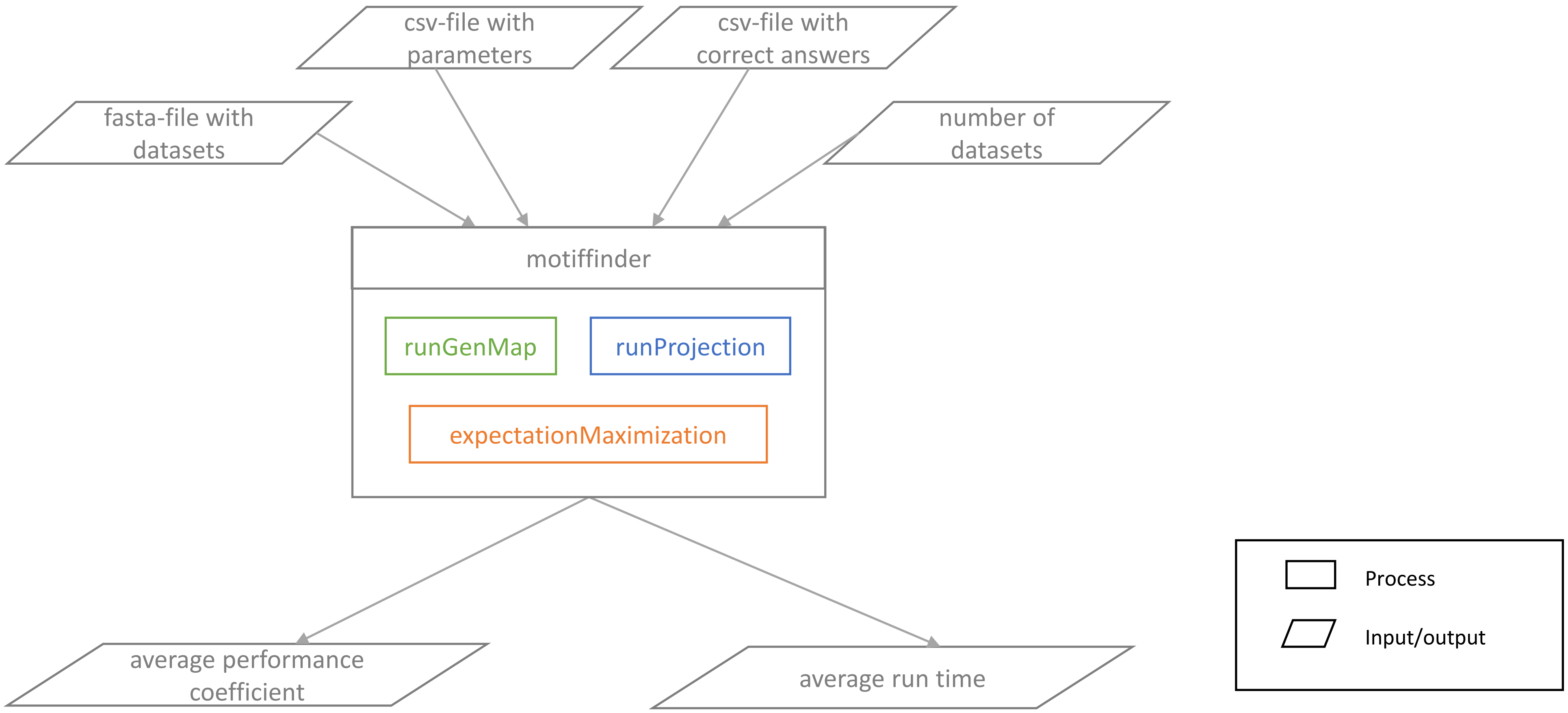
Projection

Buhler J, Tompa M. Finding motifs using random projections. J Comput Biol. 2002;9(2):225-42. doi: 10.1089/10665270252935430

Strategy:

- Assumes the mutations in a motif occurrence are uniformly distributed
- Categorizes the similarity between l-mers
- Repeats with different initializations

Program structure

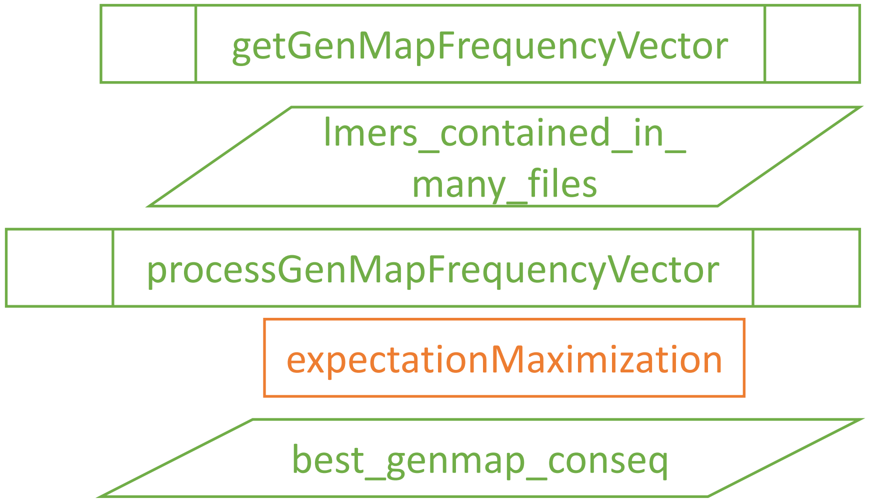


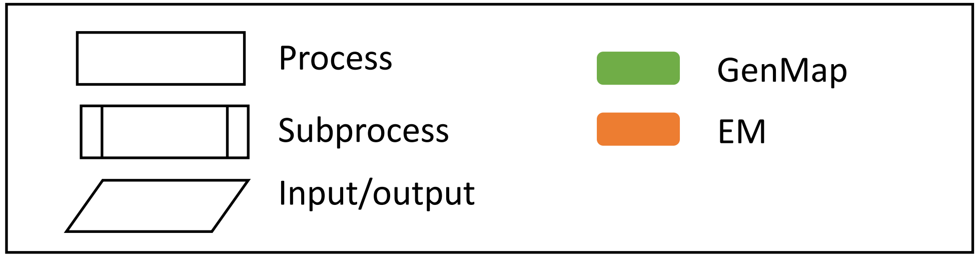
$$\frac{|K \cap P|}{|K \cup P|}$$

K is the set of the nucleotide positions of the planted motif
P is the set of the extrapolated positions by the algorithm.

Steps of runGenMap

- Generate a frequency vector using SeqAn‘s GenMap
- Filter the l-mers
- Categorize the filtered l-mers
- Refine the categorized l-mers
- Produce a consensus sequence

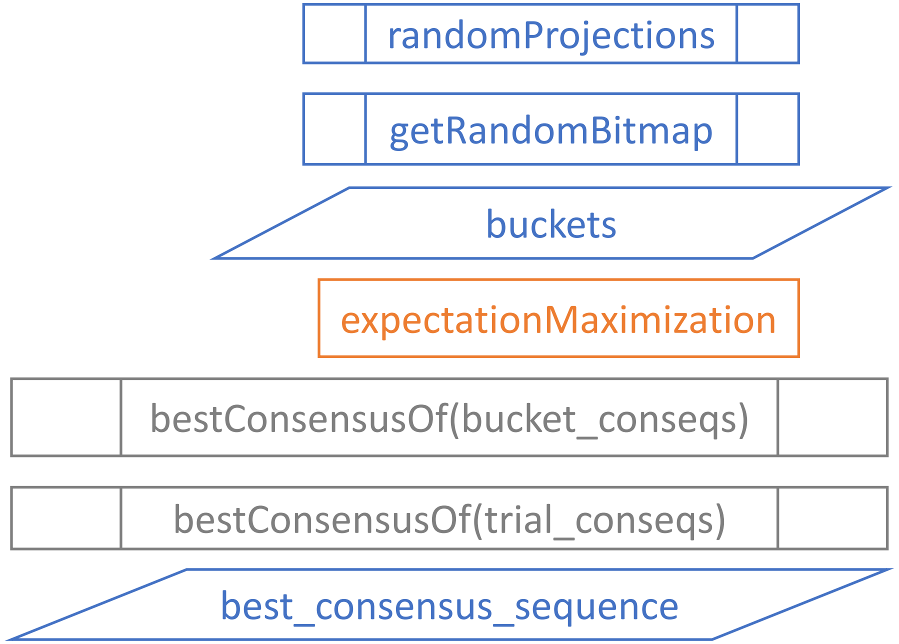


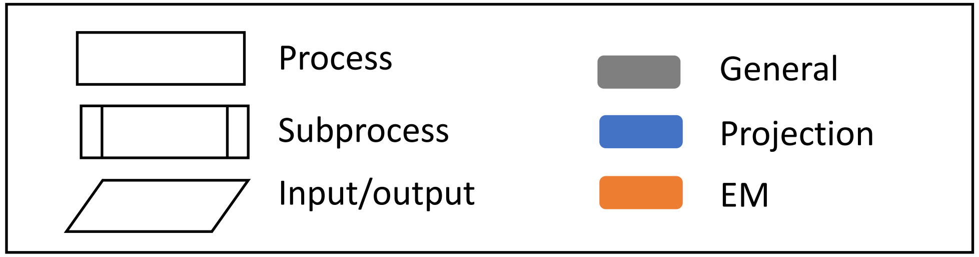


Which is the best consensus sequence?
A score is calculated which equals to the number elements in the promoters whose hamming distance to the consensus is greater than d.

Steps of runProjection

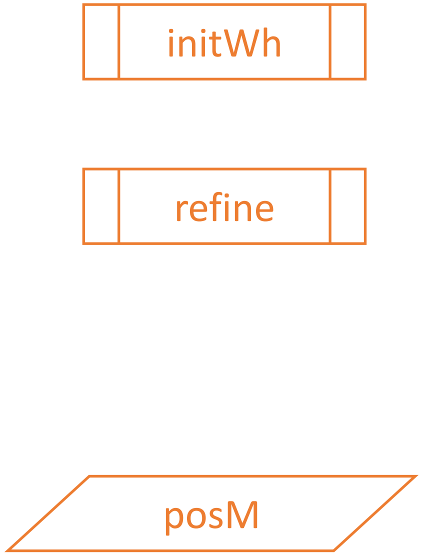
- Consider only a k number of positions in the l-mer
- Assign a hash value to the new string (projection)
- Categorize the projection into buckets
- Refine the best buckets
- Produce a consensus sequence
- Repeat m more times with different k-mers
- Take the best consensus sequence of all

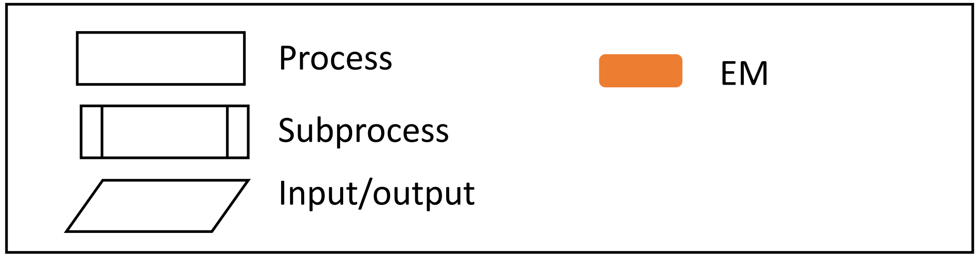




Steps of expectationMaximization

- E-Step - Initialize the weight matrix Winit
  Gives the frequency of a given base among a given position of all l-mers in the bucket
- M-Step - Calculate the position matrix
  Gives the probabilities that the motif starts at a given position in each sequence
- Iterate between the weight and the position matrix until convergence
- The refined position matrix is then used to generate a consensus sequence





Generating synthetic datasets

Creating datasets

```
/path/to/create_datasets/build ./createdatasets t n l d
```

Implanting a motif

```
/path/to/implant_motif/build ./implantmotif t n
```

where:

- t is the number of the sequences
- n is the length of one sequence
- l is the length of the motif
- d is the maximum number of mutations the motif can have

Exprimentally verified motifs

- 600 nt-long promoter regions of 16 genes in Saccharomyces cerevisiae S288C regulated by Gcn4
- In the NCBI database under “Genomic regions, transcripts, and products”
  - click on Tools -> Markers and give in the desired range
  - click on Download -> Download fasta -> Fasta all markers

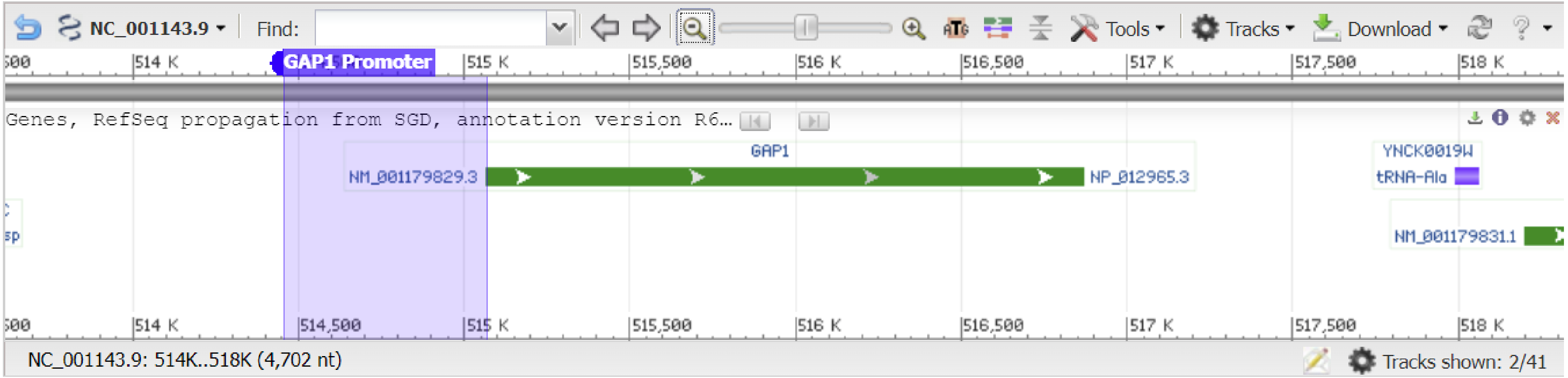


*Image source: https://www.ncbi.nlm.nih.gov/gene/853912

Running the motiffinder

For synthetic sequences

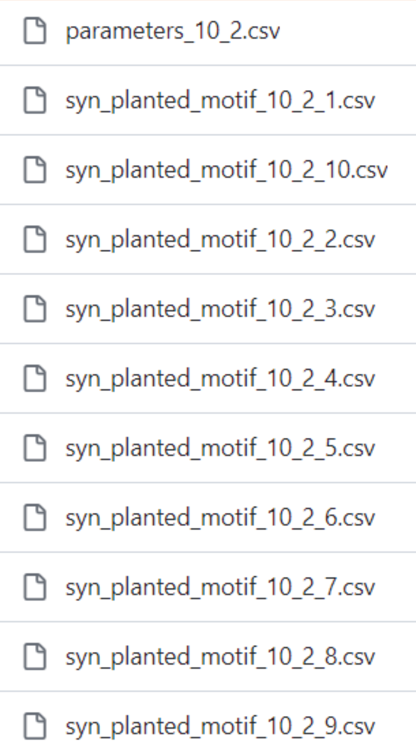
```
/path/to/build ./motiffinder ../path/to/the/synthetic/fasta/files
../path/to/the/parameters/csv/file
../path/to/the/correct/results/csv/files  numberofdatasets
```

For biological data

```
/path/to/build ./motiffinder ../path/to/the/synthetic/fasta/file
../path/to/the/parameters/csv/file
```

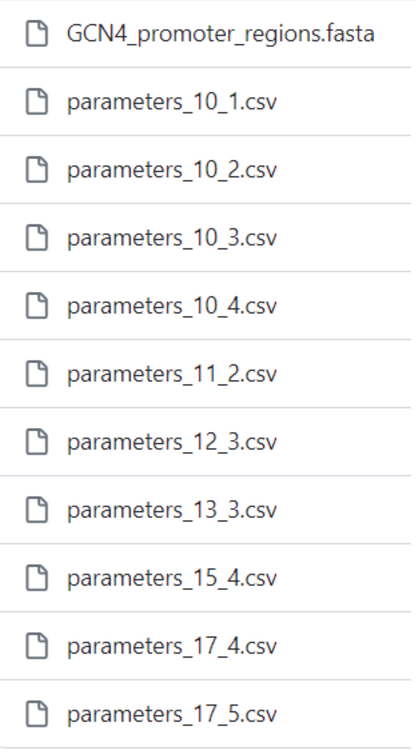
Input

Synthetic sequences with a synthetic motif



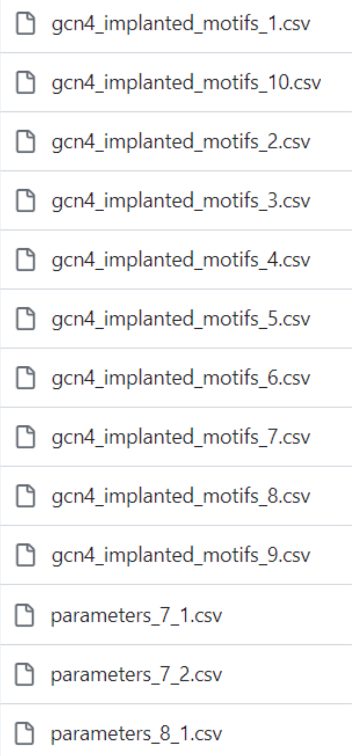
parameters_10_2.csv
syn_planted_motif_10_2_1.csv
syn_planted_motif_10_2_10.csv
syn_planted_motif_10_2_2.csv
syn_planted_motif_10_2_3.csv
syn_planted_motif_10_2_4.csv
syn_planted_motif_10_2_5.csv
syn_planted_motif_10_2_6.csv
syn_planted_motif_10_2_7.csv
syn_planted_motif_10_2_8.csv
syn_planted_motif_10_2_9.csv

Biological data



GCN4_promoter_regions.fasta
parameters_10_1.csv
parameters_10_2.csv
parameters_10_3.csv
parameters_10_4.csv
parameters_11_2.csv
parameters_12_3.csv
parameters_13_3.csv
parameters_15_4.csv
parameters_17_4.csv
parameters_17_5.csv

Synthetic sequences with a reality-based motif



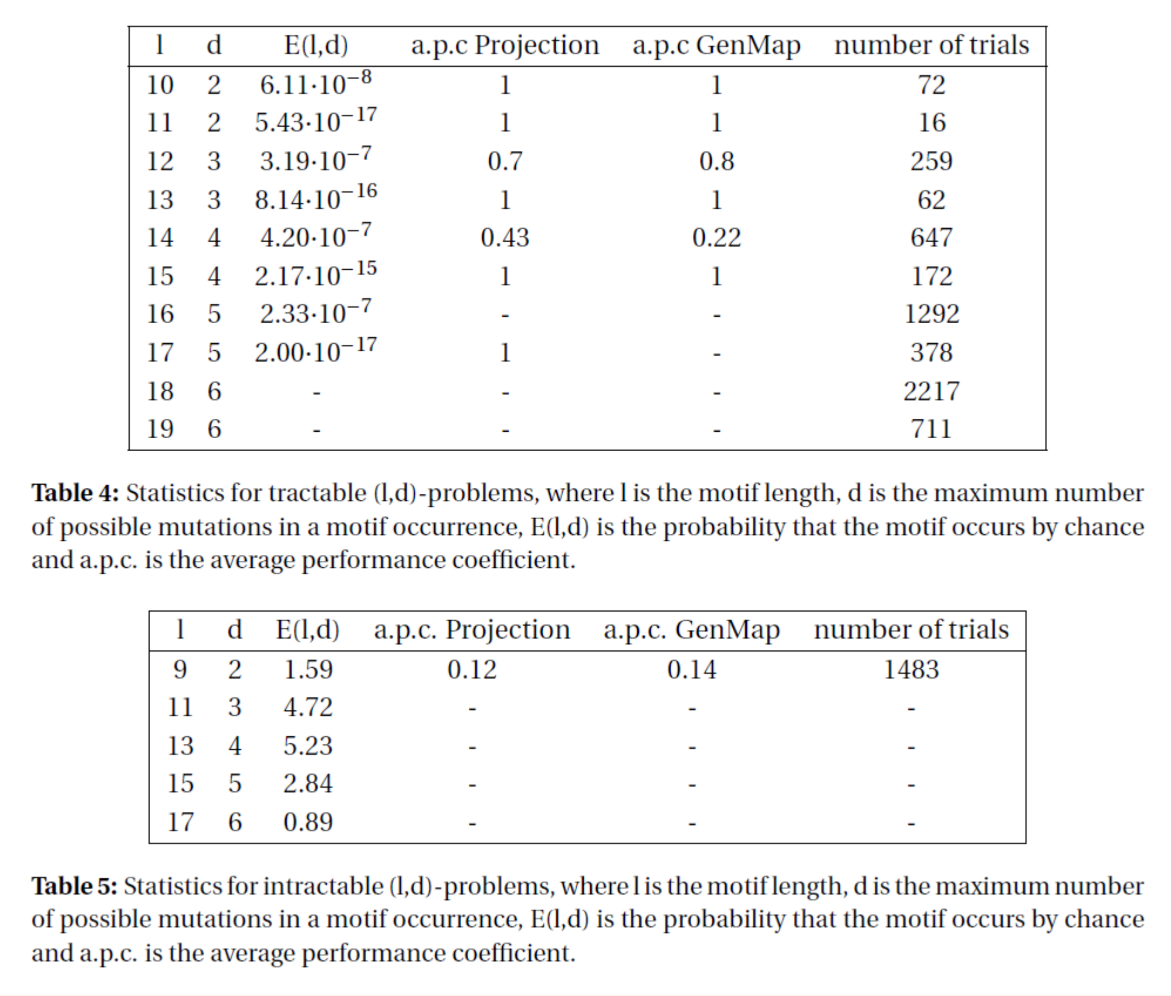
gcn4_implanted_motifs_1.csv
gcn4_implanted_motifs_10.csv
gcn4_implanted_motifs_2.csv
gcn4_implanted_motifs_3.csv
gcn4_implanted_motifs_4.csv
gcn4_implanted_motifs_5.csv
gcn4_implanted_motifs_6.csv
gcn4_implanted_motifs_7.csv
gcn4_implanted_motifs_8.csv
gcn4_implanted_motifs_9.csv
parameters_7_1.csv
parameters_7_2.csv
parameters_8_1.csv

*GitHub Repository: https://github.com/GerganaStanilova/Fast-and-exact-motif-discovery-using-the-SeqAn-library-GenMap-algorithm

Results

Accuracy

Synthetic sequences with a synthetic motif



l	d	E(l,d)	a.p.c Projection	a.p.c GenMap	number of trials
10	2	$6.11{\cdot}10^{-8}$	1	1	72
11	2	$5.43{\cdot}10^{-17}$	1	1	16
12	3	$3.19{\cdot}10^{-7}$	0.7	0.8	259
13	3	$8.14{\cdot}10^{-16}$	1	1	62
14	4	$4.20{\cdot}10^{-7}$	0.43	0.22	647
15	4	$2.17{\cdot}10^{-15}$	1	1	172
16	5	$2.33{\cdot}10^{-7}$	-	-	1292
17	5	$2.00{\cdot}10^{-17}$	1	-	378
18	6	-	-	-	2217
19	6	-	-	-	711

Table 4: Statistics for tractable (l,d)-problems, where l is the motif length, d is the maximum number of possible mutations in a motif occurrence, E(l,d) is the probability that the motif occurs by chance and a.p.c. is the average performance coefficient.

l	d	E(l,d)	a.p.c. Projection	a.p.c. GenMap	number of trials
9	2	1.59	0.12	0.14	1483
11	3	4.72	-	-	-
13	4	5.23	-	-	-
15	5	2.84	-	-	-
17	6	0.89	-	-	-

Table 5: Statistics for intractable (l,d)-problems, where l is the motif length, d is the maximum number of possible mutations in a motif occurrence, E(l,d) is the probability that the motif occurs by chance and a.p.c. is the average performance coefficient.

Results

Run time

Synthetic sequences with a synthetic motif

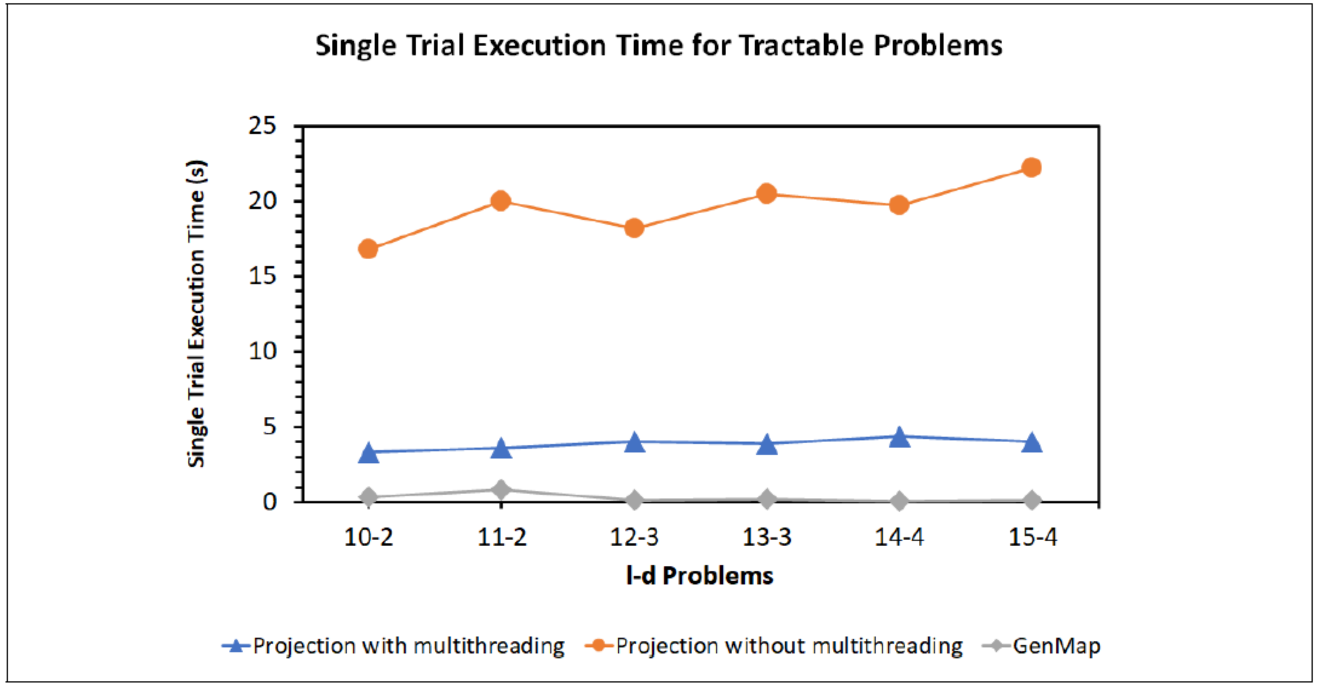


Figure 11: Line graph representing the changes in the average computational time for one trial over all 100 datasets. The (l,d) parameters are illustrated on the x-axis and the average execution time on the y-axis. The orange line shows the performance of the serial implementation of Projection, the blue line of the parallel one and the grey line of GenMap.
The chart was created using the Microsoft Excel Version 2303.

Results

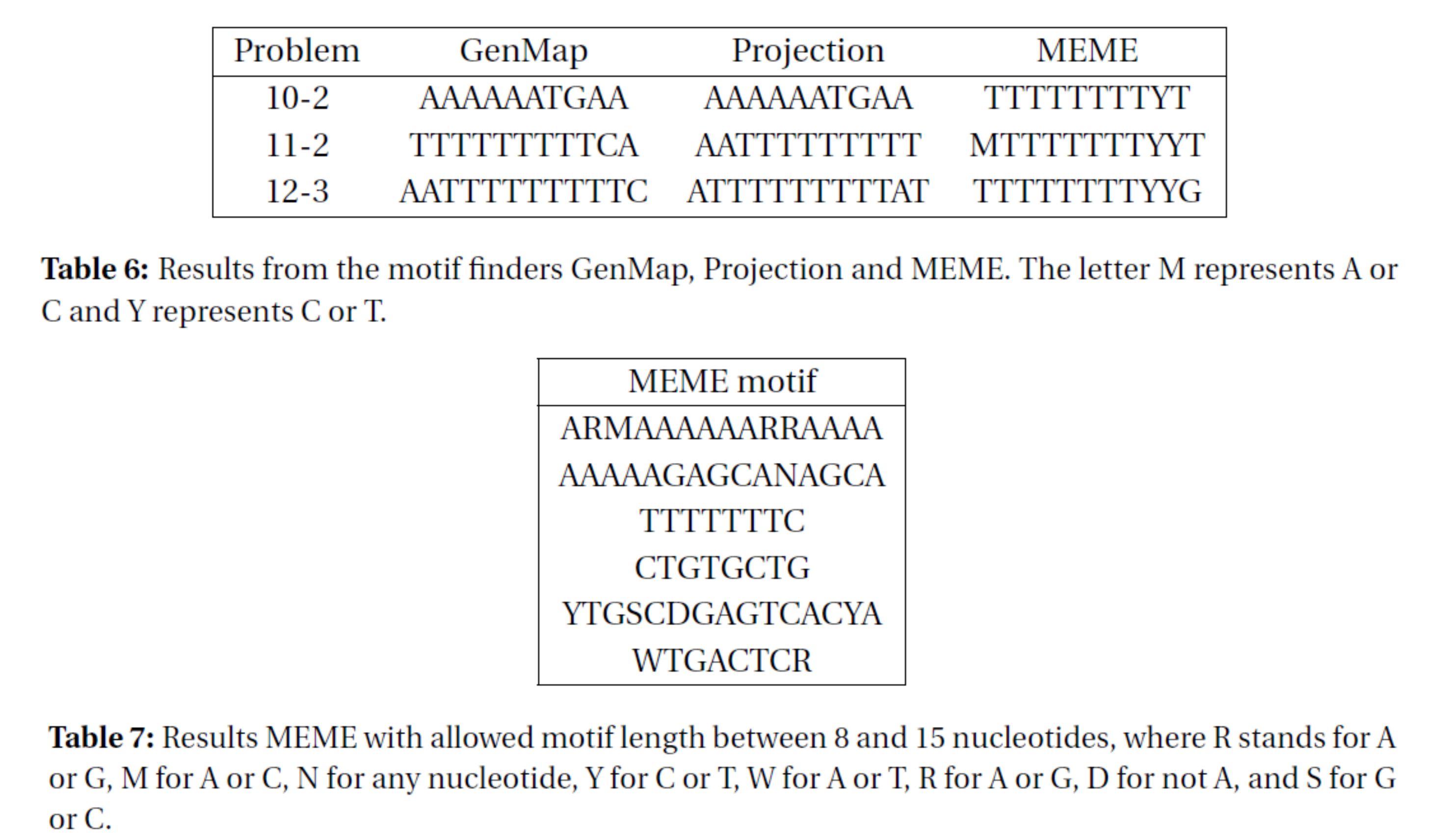
Accuracy

Consensus Sequence

of Gcn4

5' TGACTC 3'

Biological data



Problem	GenMap	Projection	MEME
10-2	AAAAAATGAA	AAAAAATGAA	TTTTTTTTYT
11-2	TTTTTTTTTCA	AATTTTTTTTT	MTTTTTTTYYT
12-3	AATTTTTTTTTC	ATTTTTTTTTAT	TTTTTTTTYYG

Table 6: Results from the motif finders GenMap, Projection and MEME. The letter M represents A or C and Y represents C or T.

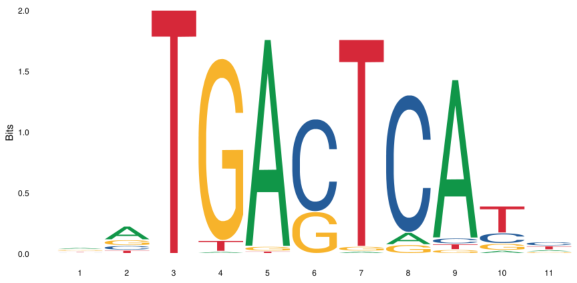
MEME motif
ARMAAAAAARRAAAA
AAAAAGAGCANAGCA
TTTTTTTC
CTGTGCTG
YTGSCDGAGTCACYA
WTGACTCR

Table 7: Results MEME with allowed motif length between 8 and 15 nucleotides, where R stands for A or G, M for A or C, N for any nucleotide, Y for C or T, W for A or T, R for A or G, D for not A, and S for G or C.

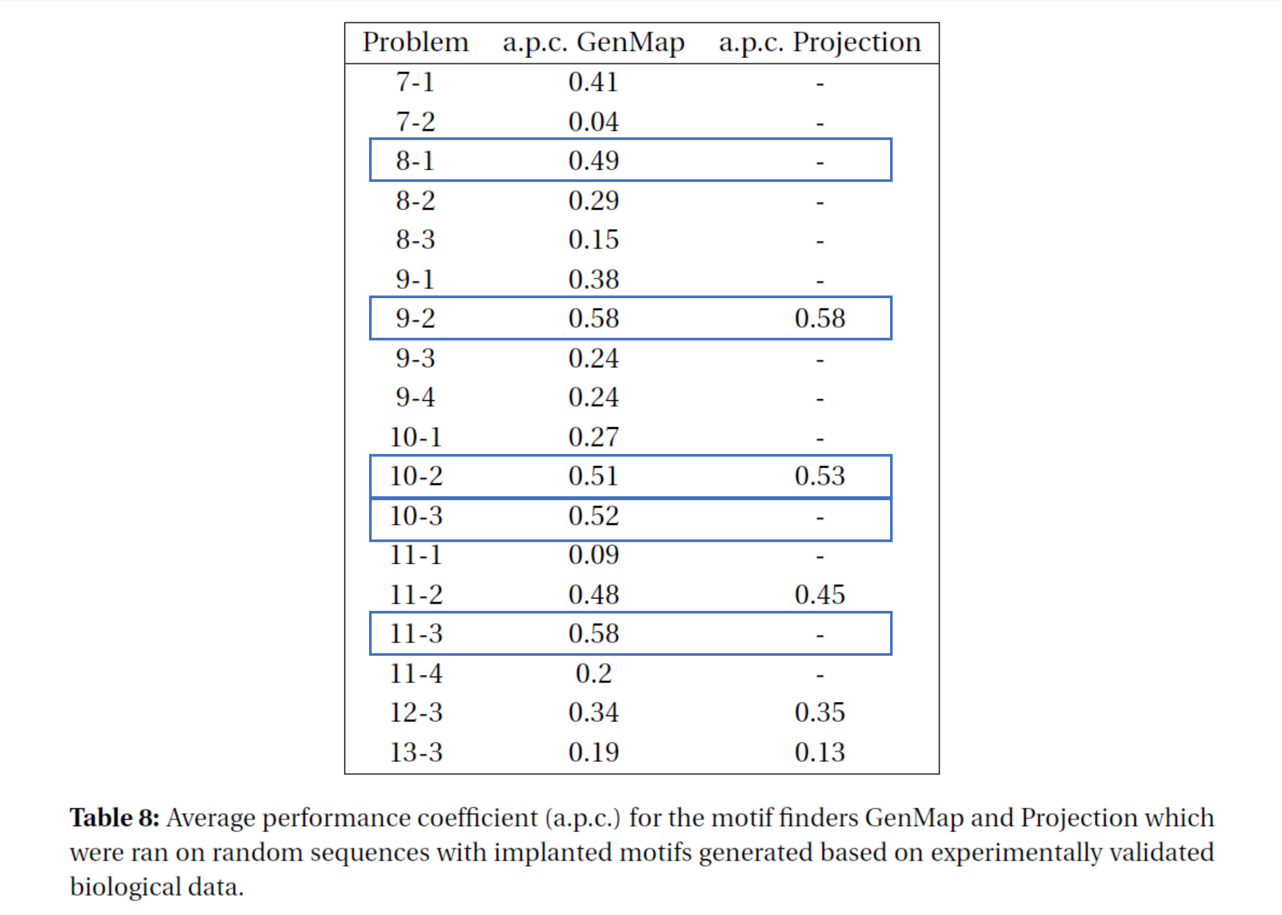
Results

Accuracy

Sequence Logo of Gcn4



Synthetic sequences with a reality-based motif



Problem	a.p.c. GenMap	a.p.c. Projection
7-1	0.41	-
7-2	0.04	-
8-1	0.49	-
8-2	0.29	-
8-3	0.15	-
9-1	0.38	-
9-2	0.58	0.58
9-3	0.24	-
9-4	0.24	-
10-1	0.27	-
10-2	0.51	0.53
10-3	0.52	-
11-1	0.09	-
11-2	0.48	0.45
11-3	0.58	-
11-4	0.2	-
12-3	0.34	0.35
13-3	0.19	0.13

Table 8: Average performance coefficient (a.p.c.) for the motif finders GenMap and Projection which were ran on random sequences with implanted motifs generated based on experimentally validated biological data.

Discussion & Conclusion

Synthetic data
Accuracy is similar but GenMap‘s run time is significantly shorter

Biological data
Suboptimal accuracy for both algorithms

Potential optimization strategies

- background single nucleotide frequencies
- larger number of datasets
- ZOOPS-model (zero or one occurrence per sequence)