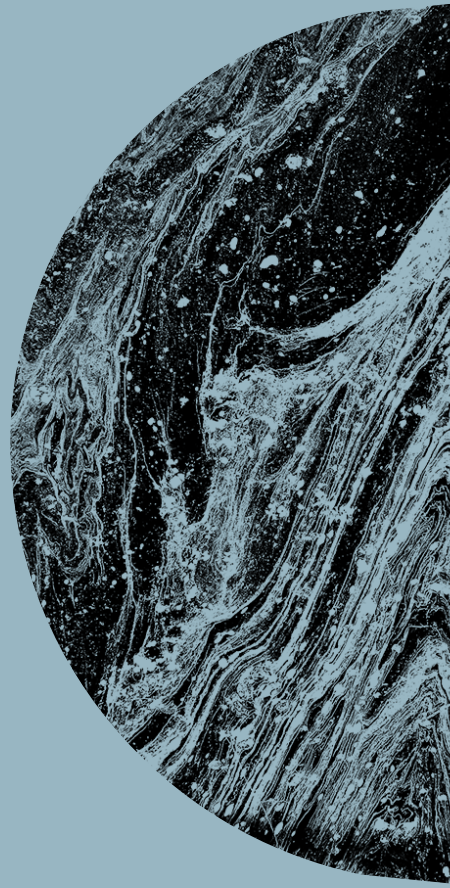
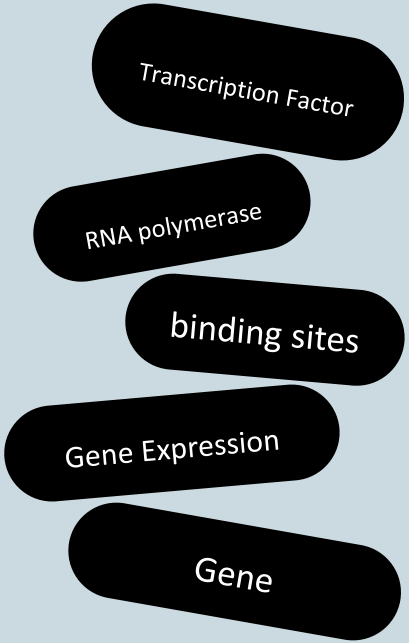
Yasamin Tafakor
Sarina Heshmati

Pseudocounts for transcription factor binding sites

Brief Explanation about Transcription Factor binding sites





gene expression is the process by which information from a gene is used in the synthesis of a functional gene product such as a protein
transcription is the process of making messenger RNA(mRNA) from a DNA template by RNA polymerase
transcription factor is a protein that binds to DNA and regulates gene expression by promoting or suppressing transcription

Introduction

Position Frequency Matrix

A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6

GAGGTAAAC

TCCGTAAGT

CAGGTTGGA

ACAGTCAGT

TAGGTCATT

TAGGTACTG

ATGGTAACT

CAGGTATAC

TGTGTGAGT

AAGGTAAGT

Position Probability Matrix

A	0.3	0.6	0.1	0	0	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0	0	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.0	0	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0	1.0	0.1	0.1	0.2	0.6

GAGGTAAAC

TCCGTAAGT

CAGGTTGGA

ACAGTCAGT

TAGGTCATT

TAGGTACTG

ATGGTAACT

CAGGTATAC

TGTGTGAGT

AAGGTAAGT

Position Weight Matrix

A	0.26	1.26	-1.32	-inf	-inf	1.26	1.49	-0.32	-1.32
C	-0.32	-0.32	-1.32	-inf	-inf	-0.32	-1.32	-1.32	-0.32
G	-1.32	-1.32	1.49	2.0	-inf	-1.32	-1.32	1.0	-1.32
T	0.68	-1.32	-1.32	-inf	2.0	-1.32	-1.32	-0.32	1.26

PWM = log (M / background)

GAGGTAAAC

TCCGTAAGT

CAGGTTGGA

ACAGTCAGT

TAGGTCATT

TAGGTACTG

ATGGTAACT

CAGGTATAC

TGTGTGAGT

AAGGTAAGT

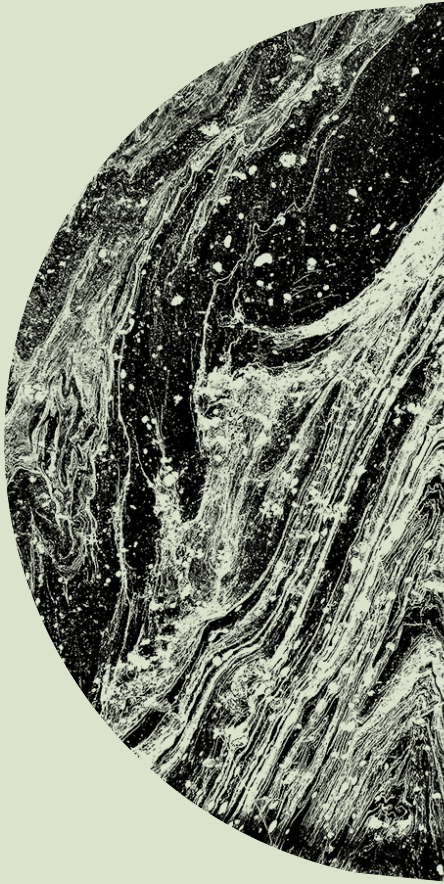
01 assigning a probability of zero is too harsh!

02 it can be troublesome to deal with negative infinity using computers.

it is common practice to add so-called pseudocounts to the PFM, in order to avoid zero probabilities.

The main approach in this study is comparison between an original PPM, which we regard as representing the real sequence specificity of each transcription factor, and a sampled PPM with pseudocounts.
The sampled motif matrix is created by stochastic generation from the original PPM. **The results show how different pseudocount choices affect the similarity of the original and sampled PPMs.**

MATERIALS AND METHODS



JASPAR dataset

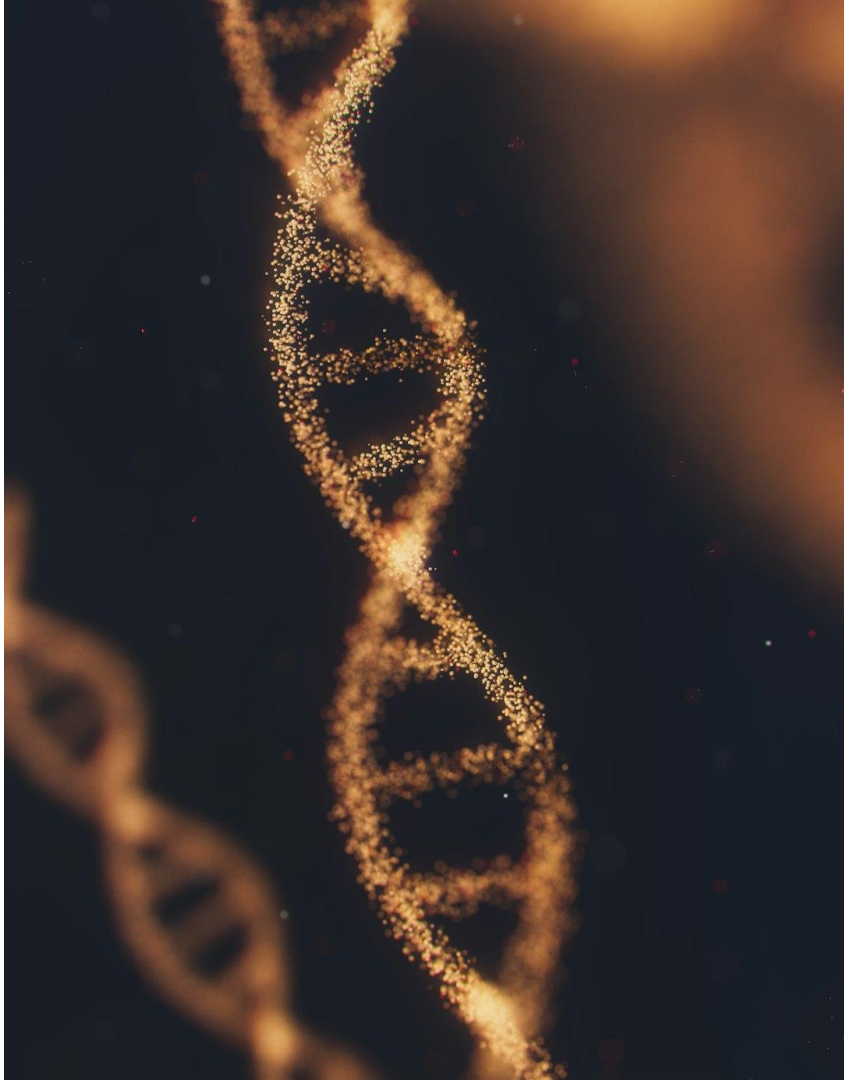
Obtained PFMs from the JASPAR database. JASPAR provides non-redundant experimentally defined transcription factor binding site motifs for multicellular eukaryotes.

Sampled PFM from original PFM

To generate sequences from a JASPAR motif, we made a PPM from the PFM.
Using these probabilities, we generated sequences randomly as virtual transcription factor binding sites. We generated sets of 10, 20, 30, 40 and 50 sequences

Pseudocount addition

For each sampled PFM, we made a sampled PPM by adding Pseudocounts.
We tried exponentially stepped pseudocounts between 0.01 and 10.
The comparison of original PPMs to sampled PPMs indicates which pseudocount is optimal.



Comparison procedures

For comparing original PPMs and sampled PPMs, we used seven methods, in two categories.
One category is matrix-based comparison, The other category is sequence based comparison

Euclidean distance

Euclidean distance (ED) was used to compare motifs

$$ED = \sqrt{\sum_{j}^{n} (s_j - s'_j)^2}$$

Total variation distance

$$TVD = \sum_{j} |s_j - s'_j|$$

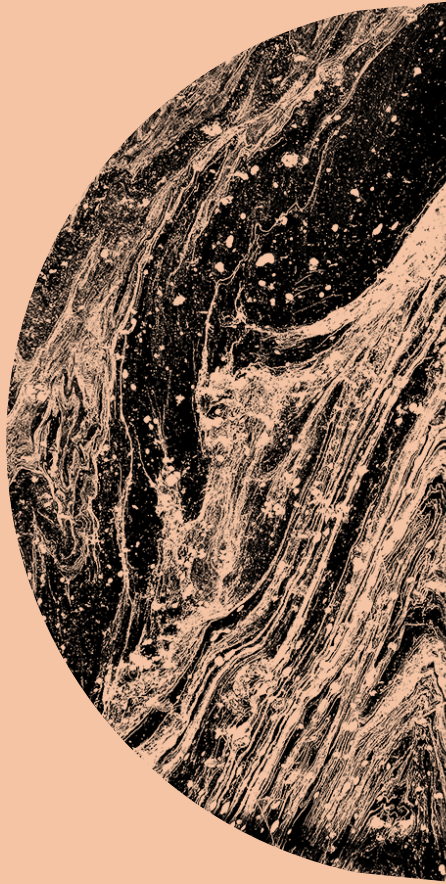
Cosine distance

$$COS = 1 - \frac{\sum_j s_j s'_j}{\sqrt{\sum_j s_j^2}\sqrt{\sum_j s_j'^2}}$$

Spearman's Rank Correlation

The probabilities of each w-mer were converted to ranks.

RESULTS AND DISCUSSION

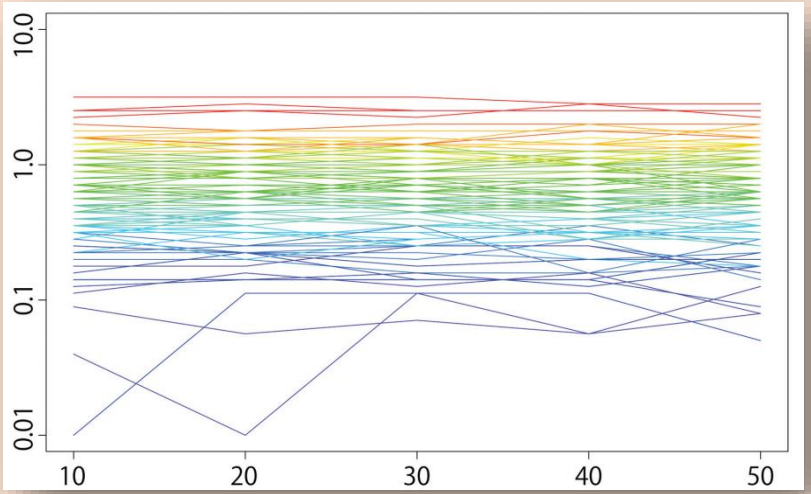


Dependence of optimal pseudocount on sample size and entropy

Each line indicates one motif, and the color represents the average entropy of the original PPM: red indicates higher entropy; blue indicates lower entropy.

Motifs with higher entropy have larger optimal pseudocounts.

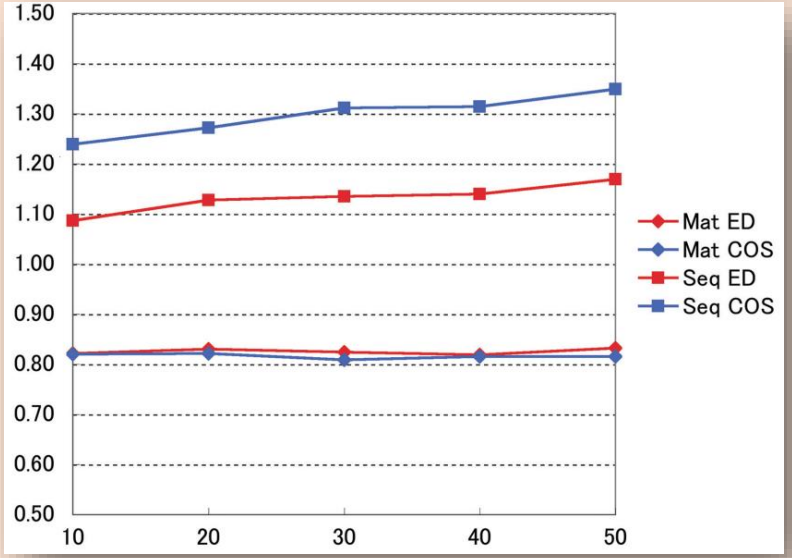
The optimal pseudocount is tightly correlated with the average entropy of the original matrix.



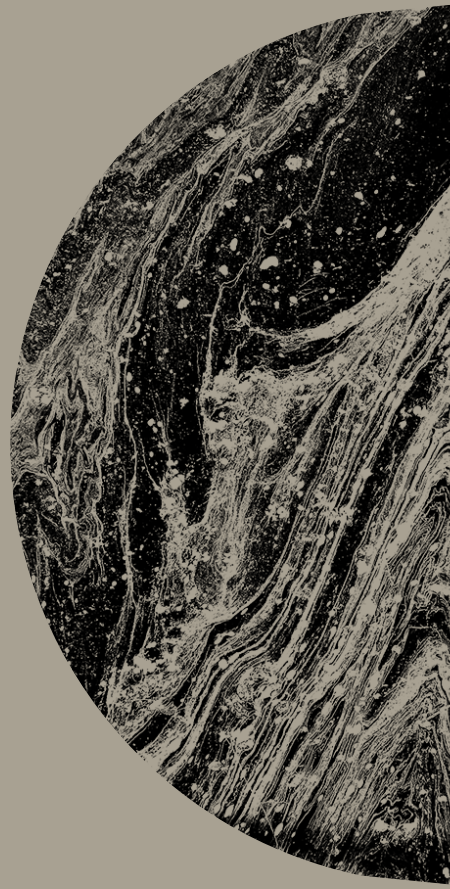
Dependence of optimal pseudocount on sample size and entropy

Perhaps surprisingly, optimal pseudocount values are not strongly influenced by sample size in most cases. Increasing sample size leads to more accurate motifs, so pseudocounts should be less necessary. On the other hand, increasing sample size reduces the impact of pseudocounts. We infer that these two effects cancel each other, when using ED.

In any case, the average optimal pseudocount over all motifs lies between 0.8 and 1.3.



CONCLUSIONS



Conclusions

First, all comparison methods indicate that pseudocounts much above 1 are a poor choice...

The results using ED and COS suggest that values close to 1 are optimal, although the optimal pseudocount depends on the entropy of the original motif.

For large sample sizes, however, pseudocounts much less than one are only marginally worse.

In summary, depending on the comparison method, optimal pseudocounts for transcription factor binding motifs are either around 1, or very low.

Thank you

yasamintafakor@gmail.com
sarinaheshmatii@gmail.com