**BACHELOR THESIS**

**Analysis of autologous binding preferences**

**of RNA-binding proteins**

submitted by

**Moritz Haderer**

in partial fulfillment of the requirements for the degree of

**Bachelor of Science (BSc)**

Vienna, 2022

**UA 033 630**

**Bachelor’s degree programme Biology**

**Univ.-Prof. Dr. Bojan Zagrovic, BA**

Degree programme code as it appears on the

student record sheet:

Degree programme as it appears on the

student record sheet:

Supervisor:

Thesis writing

1. Abstract

RBPs, complementarity hypothesis, fuzzy binding

1. Introduction

What are RBPs

what’s the complementarity hypothesis

Arthur and Thomas’ work

what is my thought behind using ppms

what datasets did I use

Approach I used (FIMO, matrices)

1. Materials and Methods  
   1. Datasets used
      1. SELEX, ATtRACT, CLIP – CDS, 5UTR, 3UTR
      2. MANE as background, Markov-model with nt-distribution
   2. Thomas‘ and Arthur’s work  
      Using exact matching, in-silico translation, results
   3. PPMs, PWMs, PFMs, PSSMs
      1. Creating PFMs, converting to PPMs, PWMs
      2. Scoring
         1. Compare matrices with exact matching
         2. PSSM & binding energy
      3. Threshold
         1. P-Value approach
         2. Q-Value?
      4. Background
         1. Markov-models higher order

How can higher order influence match-finding

* 1. FIMO
     1. What is MEME Suite; What is FIMO
     2. Statistical approach
        1. Why is the background chosen as it is
        2. Calculating p-values, scores; approximation
           1. plot: score distribution

1. Autologous Binding/Results?
   1. Analysis of FIMO results + nice plots
2. Discussion
3. Appendix
   1. code
   2. fimo commands
4. Table of contents

[2. Materials and Methods 19](#_Toc98842797)

[2.1 Datasets 19](#_Toc98842798)

[2.1.1 ATtRACT 19](#_Toc98842799)

[2.1.2 HT-SELEX 19](#_Toc98842800)

[2.1.3 MANE select transcriptome 19](#_Toc98842801)

[2.2 Previous work 19](#_Toc98842802)

[2.2.1 Exact matching 19](#_Toc98842803)

[2.2.2 In-silico translation 19](#_Toc98842804)

[2.3 My project/Generalizing the notion of “matching”/ 19](#_Toc98842805)

[2.3 FIMO 19](#_Toc98842806)

# 2. Materials and Methods

## 2.1 Datasets

### 2.1.1 ATtRACT

ATtRACT is a

### 2.1.2 HT-SELEX

### 2.1.3 MANE select transcriptome

## 2.2 Previous work

### 2.2.1 Exact matching

### 2.2.2 In-silico translation

2.2.3

## 2.3 My project/Generalizing the notion of “matching”/

**Position frequency matrix**

Experiments for finding new motifs of RNA binding proteins will yield a number of motifs that have been bound by the RBP. These motifs will often be similar in length and composition. To account for non-similarities between bound motifs, the Position Frequency Matrix helps visualize the number of times a given nucleotide was found at position P of the motif. As the name suggests, it indicates the frequency of nucleotides at a position.

A list of ten motifs bound by a given RBP in an experiment might look like this:

…CGGC**AGUAA**ACCUCA…

…UGGC**AGUAC**GCAUGA…

…AGCC**AGUAC**UCAUCC…

…UGGC**AGUAA**CCAUUU…

…GGUC**AAUAA**CCAUCA…

…UGGC**AAUAA**UCAAAG…

…CGAU**AGUAA**GCAGCA…

…CUGC**ACUAA**UUACCC…

…UGUC**AGUAA**CAAUGA…

…UGGC**AGUAA**GCAUCA…

By counting the number of times a nucleotide is observed in a given position, a PMF of the following structure can be constructed:

**Position frequency matrix (PFM)**

**A C G U**

10 0 0 0

2 1 7 0

0 0 0 10

10 0 0 0

8 2 0 0

A picture containing text, clock

Description automatically generated

Generated from motif counts in experiment

**FIND STUDIES THAT USED PFMs FOR STUFF**

Dustin E. Schones, Pavel Sumazin, Michael Q. Zhang, Similarity of position frequency matrices for transcription factor binding sites, Bioinformatics, Volume 21, Issue 3, 1 February 2005, Pages 307–313, <https://doi.org/10.1093/bioinformatics/bth480>

**Found method for comparing PFMs to each other in TFBS search**

**Position probability matrix**

Similarly to the position frequency matrix, the position probability matrix (or PPM) gives clear insight into the proportions of nucleotides bound at a given position. The main difference between the PFM and the PPM is a count-normalization to probability values (ranging from 0 to 1). This is done by dividing the frequency by the number of bound motifs. Continuing with the example from above, the corresponding PPM would be created by dividing each frequency count by ten:

**Position probability matrix (PPM)**

**A C G U**

1 0 0 0

0.2 0.1 0.7 0

0 0 0 1

1 0 0 0

0.8 0.2 0 0

Table

Description automatically generated

In practice, the position probability matrix can be used to give a probability score to a novel motif by multiplying the letters’ corresponding probabilities over all positions. In a programmatical context, one usually deals with rather long/a large number of motifs to assign a score to. Potential downsides of this type of matrix are 1. multiplication is a more expensive computation than, e. g., addition. Once the sequences to be scored exceed a certain length, this is noticeable in the duration of computations. Another downside is that multiplication of very long matrices can lead to **underflow** when computing the score of a motif. As one is multiplying many small numbers, floating point precision can be thought of as a resource. This effect might skew the scoring and, consequently, determination of sufficiency of a score.

**Position-specific scoring matrix**

To remedy both downsides of the PPM, a third type of matrix can be employed. The position-specific scoring matrix (PSSM), also called position weight matrix, can be derived from the PPM. It incorporates the background distribution of nucleotides of the target sequences. By computing the log-likelihood ratio of a nucleotide’s probability given the background distribution, the matrix’ entries can now be both positive and negative. The reason behind the alternative name “position weight matrix” is given by the interpretation of what these resulting values express: a positive value puts strong weight on a given nucleotide, while a negative value indicates, that this nucleotide should occur less often at that given position than the background distribution should suggest, thus reducing the weight.

An example calculation is easily done. First, let’s assume a uniform background distribution of 0.25 per nucleotide. Now, each value can be plugged into the equation:

Diagram

Description automatically generated

where:

Text

Description automatically generated



The log-likelihood ratio L is given by the binary logarithm of the probability of nucleotide i at position j divided by the probability of a nucleotide as given by the background distribution B.

Applying these manipulations to the PPM in the previous section, we get the following matrix:

**Position-specific scoring matrix (PSSM)**

**A C G U**

-1.32 -inf -inf -inf

-0.32 -1.32 1.49 -inf

-inf -inf -inf 2

2 -inf -inf -inf

0.8 -0.32 -inf -inf

A picture containing text, device

Description automatically generated

**Pseudocounts**

Since the logarithm is not defined at zero, the resulting matrix contains placeholders for an undefined value – often in form of negative infinity. A simple method to alleviate this issue is by adding pseudocounts to the probability values in the PPM. A pseudocount can be 0.1 or it can be calculated based on the distribution of nucleotides in the background ?

## 2.3 FIMO

**Methodology**

[1]

* Manages to give “strong” and “weak” occurrences of motifs a weight, independently of cutoffs
* “motif finding problem”: reveal statistical enrichment of certain (slightly) variable subsequences that serve as binding sites for transcription factors/RBPs?
* “counting” is a vague problem with multiple approaches:
  1. reporting whether a “match” is found or not (and how good the match is)
  2. Does the “best” match occur?
  3. sampling probability → “the probability of sampling a given k-mer from the probability distribution induced by a k-length PWM.”  
     “Naive approach: count all k-length substrings that have sampling probability above threshold”
  4. w-score: quantifies the total number of occurrences of a PWM in a sequence

Position-specific scoring matrix

<http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec285>

in biopython, this is the pwm

Biopython has a variety of methods to go from PFM to PPM to PWM to PSSM. Adding pseudocounts (ensuring log-odds ratio doesn’t turn to -inf) and a background distribution can be added if needed.

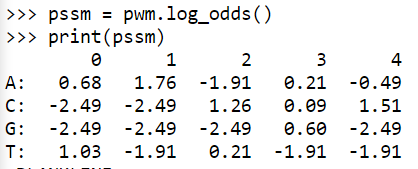
Background distributions: di-nucleotide frequency? maybe calculate distribution of letters in background transcripts to create randomized sequences from there? might eliminate bias.

Take into account negative sequences while searching for occurrences?

PSSM = log-odds ratio of a nucleotide appearing against a certain background distribution (i.e. dinucleotide frequency, etc)

Negative number: symbol appears less frequently in sequence than background would suggest

positive number: inverse



Setting a score threshold

false-positive rate: probability of “finding” a motif instance in background sequence

false-negative rate: prob of “not finding” an instance generated from the motif

ROC curve?

Patser threshold by Hertz and Stormo [2]

[3] Sensitivity and Specificity + ROC curve for threshold determination:

sensitivity: probability that a match is actually a match

specificity: probability that a non-match is actually a non-match

1-specificity: false positive rate: probability that a match is actually a non-match

ROC-curve: 1-specificity vs sensitivity

Scoring a matrix

Weight matrix values are very small (usually 0.x) and we risk underflow

Remedy: log-likelihoods are used; they allow us to add values instead of multiplying. Higher values give more insight because they differentiate different matches more clearly.

Log-likelihood ratio = Log(observed frequency / expected frequency)

Information content measured in bits

Including nucleotide dependency: the crux of PWMs

Higher order PWMs

Maximal Dependence Decomposition (MDD)

[1] On counting position weight matrix matches in a sequence, with application to discriminative motif finding; Saurabh Sinha

[2] Identifying DNA and protein patterns with statistically significant alignments of multiple sequences; G Z Hertz 1, G D Stormo

[3] Master’s thesis: “Higher order PWM for Modeling TF binding sites” (by Dhivya Srinivasan; in “papers” folder on lab pc)

Autologous binding

FIMO

Why FIMO rather than MAST?

* MAST reports a single score for each sequence - not one score per motif;
* MAST is useful when you need to rank your sequences based on how well each sequence matches all of your motifs

Inputs

**“Motif file”**

must contain “meme formatted motifs” for which there are different conversion tools available.

Our data:

* ATtRACT\_ppm:
  + matrix\_id associated with a certain RBP in the ATtRACT\_db file
  + length of the motif
  + PPM
* SELEX2020\_ppm:
  + Protein identifier (according to ENSEMBL)
  + motif length
  + gene name
  + mono/multimer status
  + PPM
* CLIP data
  + In FASTA format
  + nonamer? hexamer?

**“Sequence file”**

file containing sequences the motifs are scored against; in FASTA format

Processing Details

fimo -options motif-file sequence-file

-options:

-bfile - includes a background

Alphabet: ACGU

**Background**

Can be provided together with motif-file in command-line argument. They tell the MEME Suite how prevalent each letter of the motif alphabet was in the source sequences that were used to create the motifs.

Using FIMO via the web

* Can use a sequence database that’s provided; FIMO will use a background model based on the sequences provided

Using FIMO via command-line

* fasta-get-markov command creates a 0-order markov model from fasta sequence file
  + I used the command on the MANE.fa file and got a background file; the MANE sequences included all CDS, 3UTRs and 5UTRs (i think)

**Scoring**

Each input motif is converted to a PSSM (using log-odds)

What counts as an occurrence?

* “statistically significant log-odds score”
* p-value (so what is deemed “significant”) can be controlled

**Threshold setting**

Uses a dynamic programming algorithm to convert log-odds scores into p-values, assuming i.i.d. background. Program reports occurrences with a p-value of less than 1e-4.

Outputs

<https://meme-suite.org/meme/doc/fimo-output-format.html>

Next Todos

My analysis doesn’t yield the results I expected

* What difference do pseudocounts make? A PWM with some zeros would turn to -inf, how does FIMO handle that?
* change p-value cutoff for FIMO
* put one sequence into arthur’s code and into fimo and compare results

Troubleshooting

11.3.

absolutely no enrichment in z-values of autologous matches vs. all matches

MEAN-calculations: Do I calculate the mean over ALL matches in, e.g., the CDS? Or do I calculate it per motif? So the mean coverage of all sequences that matched with a given motif. Motif A had matches on 100 sequences; 100 different coverage values. Among those 100 there’s the autologous sequence

P-values and backgrounds:

graph shows skewed distribution?

backgrounds are plotted?

how do I analyze it independently of arthur’s?

plot background distributions?

It definitely has to do with the background to every autologous match

an autologous match happens in an environment of many matches of a motif with other sequences. The motif has a certain coverage value on each sequence and I want to know whether the motif is more likely to bind to its own mRNA than it is to bind to any other mRNA.

As far as I can tell in my analysis, this is not the case. However, we know that it IS the case.

**How is the p-value calculated? What does each line in the plot say?**

Notes on Arthur’s code

* function: randomized\_nucleotide\_PWM\_search  
  every autologous score (ppm of an RBP is run over that RBP’s mRNA sequence) is calculated.   
  Then, the autologous mRNA sequence is put into a random-sequence-generator function, which fishes sequences from the MANE transcriptome database that are longer or as long as the autologous sequence. The nucleotides of those sequences are then randomly shuffled around so as to make them random.  
  That means: every autologous coverage value is of the RBP motif against its own mRNA; Every “background” coverage value is of the same RBP against a number of randomly generated sequences.

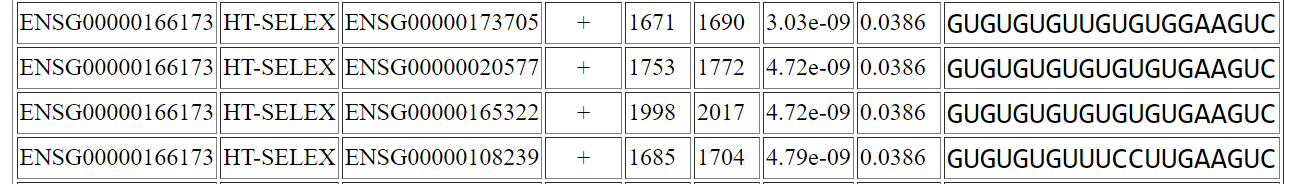
The research question therefore is:   
**Does an RBP’s motif bind its own mRNA with higher affinity than it binds any sequence of randomly shuffled nucleotides with a nt-frequency of a given transcriptome sequence of equal or higher length?**

What is my research question?

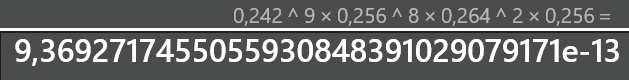
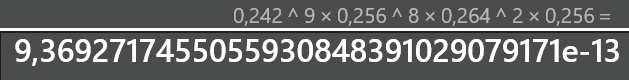
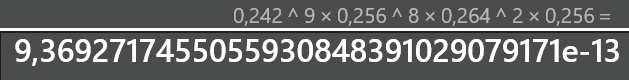
**Does an RBP’s motif bind its own mRNA with higher affinity than it binds other mRNAs in the human transcript?**

**Find out what FIMO does:**

**Various matches:**

****

**p-values for these matches:**

1. **9xG, 8xU, 2xA, 1xC = **
2. **9xG, 8xU, 2xA, 1xC**
3. ****
4. **9xG, 8xU, 2xA, 1xC**
5. ****
6. **6xG, 9xU, 2xA, 3xC**

**Background:**

****

**Best match:**

****

Best match: 7xG, 7xU, 4xC, 2xA

**FIMO paper:**

Charles E. Grant, Timothy L. Bailey, and William Stafford Noble, "FIMO: Scanning for occurrences of a given motif", *Bioinformatics*, **27**(7):1017-1018, 2011.

# Enables running MEME Suite

export PATH=$HOME/meme/bin:$HOME/meme/libexec/meme-5.4.1:$PATH

**command: bg=transcriptome; cutoff=5e-2**

fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/selex\_UTR5 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/selex\_CDS --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/selex\_full --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/selex\_UTR3 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/rnacomp\_UTR5 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/rnacomp\_CDS --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/rnacomp\_full --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/rnacomp\_UTR3 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt  FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/htselex\_UTR5 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/htselex\_CDS --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/htselex\_full --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval5e-2/htselex\_UTR3 --max-stored-scores 2147483646 --thresh 5e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt

**Command: background:full transcript; Threshold: 1e-2**

fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/selex\_UTR5 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/selex\_CDS --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/selex\_full --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/selex\_UTR3 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/rnacomp\_UTR5 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/rnacomp\_CDS --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/rnacomp\_full --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/rnacomp\_UTR3 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt  FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/htselex\_UTR5 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/htselex\_CDS --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/htselex\_full --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-2/htselex\_UTR3 --max-stored-scores 2147483646 --thresh 1e-2 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt

**command: bg=transcriptome; cutoff=1e-3**

fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/selex\_UTR5 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/selex\_CDS --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/selex\_full --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/selex\_UTR3 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/rnacomp\_UTR5 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/rnacomp\_CDS --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/rnacomp\_full --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/rnacomp\_UTR3 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt  FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/htselex\_UTR5 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/htselex\_CDS --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/htselex\_full --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-3/htselex\_UTR3 --max-stored-scores 2147483646 --thresh 1e-3 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt

**command: bg=transcriptome; cutoff=1e-4**

fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/selex\_UTR5 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/selex\_CDS --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/selex\_full --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/selex\_UTR3 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/rnacomp\_UTR5 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/rnacomp\_CDS --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/rnacomp\_full --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/rnacomp\_UTR3 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt  FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/htselex\_UTR5 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/htselex\_CDS --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/htselex\_full --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt;fimo --oc FIMO\_OUT/background\_transcriptome\_pval1e-4/htselex\_UTR3 --max-stored-scores 2147483646 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt

Take a step back.

Basic questions:

**Do RBPs bind their own mRNAs more than the average mRNA?**

What did Arthur do?

He took the ppms and let them slide over the sequence. Whenever a score was high enough (above threshold), he made the zeroes in his sequence into ones.

Then, he calculated nt\_nt\_motif by getting the mean of the array.

In one variation, he allowed the value to go above 1; This puts emphasis on overlapping matches.

What he output was the nt\_nt\_motif of a single sequence.

For my analysis, I can take my sequences, record their matches (autologous and non-autologous) and the start+end of the match.

Then I transform the start+end into indices for each sequence and make zeros to ones whenever a match occurs. Then, for each sequence, I get the mean and thus the nt\_nt\_motif.

function analyze\_found\_motifs uses the output of the function nucleotide\_PPM\_search.

What does nucleotide\_PPM\_search do?

it takes a dictionary (attract\_ppms, htselex\_ppms) of the structure dict = {experiment:{sequence\_id:matrix\_id}}

matrix\_id is used to get the ppm from a different variable (unnecessary step). Also, a deviation (score threshold) is calculated → not necessary for me

Then, he goes through all cDNAs in MANE (the original dict\_MANE) and runs the ppm over the sequence using **perform\_PPM\_slow\_sliding**. Each resulting nt\_nt\_motif score (coverage of sequence by all motifs) is saved in variable found\_motifs.

found\_motifs = {experiment:{sequence\_id:{subsequence:[nt\_nt\_motifs]}}}

**From FIMO output to plot:**

* take the tsv files and get the data into proper shape.

I want to build up the found\_motifs dictionary

**Autologous binding and how to analyze it**

RNA-binding proteins carry out a multitude of functions in the cell. It is their interactions with mRNA, the carrier sequence for genetic information, that makes them so important. Despite their importance, there exist only a handful of datasets pinpointing the mRNA sequence profiles those RNA-binding proteins (RBPs) like to bind to. In a set of experiments, around 170 such binding-profiles, or motifs, were discovered and reconfirmed.

The complementarity hypothesis states that a protein will have a larger affinity towards the coding sequence (CDS) of its own mRNA. Applying this hypothesis to the world of RBPs, a set of proteins that are known to bind mRNA in the first place, we would like to affirm that an RBP has a higher propensity to bind to its own coding region than to any other mRNA sequence and subsequence.

The MANE transcriptome dataset serves as a control dataset for our analysis. It’s a dataset of 200k~ mRNAs (the whole human transcriptome), including a number of RNA binding proteins.

In order to test the complementarity hypothesis, the RNA binding motifs of our RBPs are checked against every single sequence in the MANE transcriptome dataset.

A motif consists of a probability value that any of the four nucleotides would appear at a given position in the motif. When a sequence is similar enough to the motif, then the added up probabilities surpass a set threshold score (usually set via p-values and including a background distribution of nucleotides in the sequence considered in order to eliminate bias). According to the complementarity hypothesis, the score of a given RNA binding protein’s motif when it is run over its own mRNA should be significantly higher than the average score it would reach when running over any other mRNA. Additionally, scoring the coding sequence of its own (autologous) mRNA would reach a higher score than any other subsequences.

In practice, every sequence in the MANE transcriptome is handed to every single RNA binding motif and a score is calculated.

FIMO decides the cutoff score depending on the likelihood of getting the scanned subsequence with a given distribution of nucleotides.

The files FIMO outputs give the exact position and score of a motif that matched a sequence. In theory, following the complementarity hypothesis, a motif would match its own mRNA more often than a motif would match any other sequence.

The importance of background:

In Arthur’s project, he had different ways of making sure that motif-sequence matches weren’t just caused by random chance. Mainly, he did this by randomly shuffling around the nucleotides in his sequences. That way, he has a random dataset and a non-random dataset. Their differences will then make clear, which tendencies RNA binding proteins follow when binding motifs.

In my analysis, I had no shuffling of sequences to create a randomized background. Rather, I used FIMO to create a background distribution

**On the issue of full-transcript matching vs. UTR5 matching**

A number of motifs showed the following particularity: When the motif was scored against the full transcript sequence, no matches were recorded. When it was scored against, for example, the 5’ UTR subsequence, which is, obviously, contained within the full transcript, then matches were recorded.

What could be the reason for this on the FIMO side:

What could be the reason for this on the code side:

* Wrong sorting of sequence matches to motif IDs:
  + Confused listings of sequence IDs with motif IDs?

**On the issue of results**

Since we know the results should go in one direction, I’m disappointed to see that my results show a different picture. This could either be a consequence of FIMO’s analysis of the transcriptome, but it could also be a flaw in my code.

Code side: None, everything’s checked.

I calculate the coverage of each motif-sequence pair over all matches, I get the mean and std of coverages per motif, I calculate the z-scores appropriately, create the right background and my results show that the probability of getting a given z-score of an autologous match is 1, given the background distribution

Low-complexity regions: Long repeats of a single nucleotide or a repeating stretch of two nucleotides.

Since our background distribution is nearly uniformly distributed, the p-value of getting

**Arthur’s approach**

* given a motif in the form of a position probability matrix, he used the function “approximate cutoff” to find a statistically significant score for a motif-sequence match. The function does this by calculating the score for every single subsequence that has the length of the motif, and determining the minimum score that should be valid given some threshold. The threshold is decided arbitrarily and is probably something like 0.05 (pvalue-type of threshold)
* The function perform\_PWM\_slow\_sliding does the scoring of each PPM against each sequence. Before scoring, every sequence is made into an array of zeros. Every motif-length subsequence in the human transcriptome is scored. The score gets compared to the previously decided cutoff and, if valid, the zero in the array is turned into a one. In the end, every valid match will have a 1 in the array. Then, the mean is taken, resulting in a coverage value for the sequence-motif pair
* How he handled background:
  + random di-nucleotide sequence function generates a sequence of certain length by using the di-nucleotide frequency distribution of the (transcriptome? human genome?)
  + get randomized sequence list: function that gives a certain amount of sequences (seq\_amount) of certain length (seq) and of certain “key” (e.g. which subsequence to use: UTR3, UTR5, CDS…). Randomly shuffled sequences are made.
* If there really was enrichment in autologous binding, a motif should have “stronger” matches in its own mRNA. This “strength” of matches is not even a part of the picture here. All we count is the “frequency of matches over a sequence by a motif that have a probability of 0,1% of occurring randomly”.

Theoretical differences between what Arthur did and what I did:

cutoff for what counts as a “match”.

he scored “most” sequences and determined a “min” score for that particular motif that would count as a match.

FIMO: given a certain distribution of nucleotides, a log-likelihood matrix is constructed.