**BACHELOR THESIS**

**Analysis of autologous binding preferences**

**of RNA-binding proteins**

**RNA binding motifs reveal tendencies towards autologous binding**

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:

**Abstract**

The central dogma of molecular biology [1] is a proposed model for the passing down of genetic information happening in the cell. While it captures the essence of life on a molecular level, there are types of interactions that are difficult to understand when viewed from this perspective. Interactions between RNA-binding proteins (RBPs) and RNA count among them. They are integral parts of many cellular processes such as post-transcriptional regulation, RNA transport and localization and post-transcriptional modification of RNA. As such, the levels of RBPs need to be tightly regulated to ensure proper cellular functioning. Autoregulatory feedback between RBPs and their own mRNAs [2] often lie at the heart of these regulations. The recently brought forward complementarity hypothesis helps understand how such interactions between proteins and their own mRNA can occur. As a relatively new idea, the strength of this hypothesis suffers from a lack of experimental results. Given some new data in form of protein-RNA interaction profiles, I here show a possible systematic analysis of the theoretical binding behavior of RNA binding proteins. The results of the analysis show a general trend towards autologous interactions of RBPs with their own mRNAs. These interactions happen more readily when an RBP’s binding preferences are relatively unspecific. While protein-RNA interactions can happen highly specifically, these vaguer interaction profiles point towards the alignment of certain more general physicochemical properties of these biomolecules. As the central dogma states, information flows from RNA to protein. Perhaps, this flow of information includes properties that allow a protein to recognize its original mRNA in a global manner and, therefore, bind and regulate it.

Contents

[1. Introduction 5](#_Toc100003479)

[2. Materials and Methods 7](#_Toc100003480)

[2.1. Datasets 7](#_Toc100003481)

[2.1.1. ATtRACT (2016) [1] 7](#_Toc100003482)

[2.1.2. HT-SELEX (2020) [2] 8](#_Toc100003483)

[2.1.3. Experimental techniques for motif discovery 8](#_Toc100003484)

[2.1.4. MANE select (v0.95) transcriptome 9](#_Toc100003485)

[2.2. Previous work 9](#_Toc100003486)

[2.2.1. The datasets 10](#_Toc100003487)

[2.2.2. Background 10](#_Toc100003488)

[2.2.3. Exact matching 10](#_Toc100003489)

[2.2.4. In-silico translation 11](#_Toc100003490)

[2.2.5. Position probability matrices 11](#_Toc100003491)

[2.2.6. Results 12](#_Toc100003492)

[2.3. My project/Generalizing the notion of “matching” 13](#_Toc100003493)

[2.3.1. Position frequency matrix 13](#_Toc100003494)

[2.3.2. Position probability matrix 14](#_Toc100003495)

[2.3.3. Position-specific scoring matrix 14](#_Toc100003496)

[2.3.4. Scoring an occurrence 16](#_Toc100003497)

[2.3.5. Threshold setting 16](#_Toc100003498)

[2.3.6. Background distributions 17](#_Toc100003499)

[2.4. FIMO 18](#_Toc100003500)

[2.4.1. The MEME Suite 18](#_Toc100003501)

[2.4.2. What is FIMO 19](#_Toc100003502)

[2.4.3. FIMO Input preparation 19](#_Toc100003503)

[2.4.4. Statistical background 20](#_Toc100003504)

[2.4.5. Outputs 22](#_Toc100003505)

[3. Results 24](#_Toc100003506)

[4. Discussion 27](#_Toc100003507)

[5. References 28](#_Toc100003508)

[6. Acknowledgements 30](#_Toc100003509)

[7. Appendix 31](#_Toc100003510)

# Introduction

RNA binding proteins (RBPs) are a class of proteins that interact directly with RNA molecules [3]. These interactions have been found to involve very important cellular processes such as post-transcriptional modification and regulation of translation [4], RNA transport and localization as well as mRNA splicing [5]. Although being central to many cellular processes, the interactions between these important proteins and RNAs have not been studied to the extent they merit.

During evolution, RBPs got more and more specialized to recognize specific RNAs [5], developing binding regions that show affinity only to a specific type of recognition pattern of ribonucleic acids. These patterns are often called “motifs” and are one of the central topics of this thesis.

Recognition motifs could be structurally traced back to certain binding domains that exist within the 3D-structures of proteins. These domains have developed affinity for relatively specific sequences of ribonucleotides. While some domains prefer specific sequences of nucleotides, others bind to the phosphate backbone of RNA, being less nucleotide specific. Some domains bind linear RNA molecules, others bind complex structures or even unstructured RNAs. To have a more concrete idea of protein-RNA interactions, some knowledge of the most common RNA-binding domains is useful. Among those common domains are the RNA-Recognition Motif (RRM), the Double-stranded RNA-binding Motif (dsRBM), K-Homology domains, and RG/G repeats.

The RNA-Recognition Motif is found in 0.5%-1% of human genes and consists of a stretch of 90 amino acids. With conserved regions containing a handful of aromatic amino acids that bind the RNA in a base-unspecific manner, the composition of the surrounding β-sheets allows for nucleotide-specific binding.

As the name suggests, the Double-stranded RNA-binding motif prefers to bind double-stranded RNAs. This also implies that these interactions are less specific, as the base-information is not as readily available as on a single-stranded nucleotide and the domain must resort to binding the phosphate-backbone. What these interactions lack in nucleotide-specificity, however, they make up in recognition of complex-structured RNA-molecules.

K-Homology domains are commonly able to bind a wide variety of four-nucleotide long sequences. As this alone would not suffice in efficient motif recognition, this domain has found ways of increasing its specificity. In some situations, a neighboring α-helix is integrated into its RNA recognition domain, thereby allowing efficient binding of six nucleotides. In other cases, KH-domains appear in repeats, which further increases specificity.

RG/G (Arginine-Glycine/Glycine) repeats are RNA-recognition domains that can interact with RNA molecules via their Arginine residue. They are usually found in intrinsically disordered protein regions and are hypothesized to play a role in protein phase-separation. As a single RGG region does not show enough specificity for efficient RNA binding, they are most often found in repeats, thereby increasing specificity.

The recently proposed “complementarity hypothesis” creates a framework for understanding these protein-RNA interactions in a general sense. According to this framework, whenever genetic information flows from mRNA to protein, the physicochemical properties of the mRNA sequence are translated to the physicochemical properties of the protein. Hence, the stretch of mRNA that codes for a given RNA-binding domain transfers a certain affinity towards itself. The complementarity hypothesis therefore states that RNA-binding domains have general affinity towards the mRNA they originated from. This flow of information is what constitutes the genetic code, being a dictionary for mRNA-to-protein translation. In fact, it has been found that translation does not simply go from base to amino acid, but rather involves passing down physicochemical properties. This phenomenon yields fascinating insight into affinity of proteins to their cognate mRNAs [6].

As experimental results for the complementarity hypothesis are scarce, we resort to in-silico analysis of protein-RNA interaction data for insight. The emergence of high-precision technologies like RNAcompete and SELEX, being methods that enable the determination of the exact nucleotide motifs a given RBP can recognize, facilitates this type of analysis a great deal. RNAcompete and SELEX can determine RBP motifs by incubating RBPs with a large variety of RNAs. After incubation, RBPs which bound RNA are eluted and RNA-sequencing is used to determine exact binding motifs [7] [8].

### Experimental techniques for motif discovery

Diagram

Description automatically generated

Figure 1: Description of experimental techniques for RBP binding preference determination [12]

**RNAcompete** [7] is an in-vitro experimental technique where a Glutathione-S-Transferase-tag is added to an RBP of interest. The modified RBP is then incubated in a large pool of RNAs (~ 250 000 molecules) that has been designed to cover all combinatorically possible sequences of up to 9 bases. After incubation of the RBP in an environment holding an excess of RNA, a pull-down assay is performed, eluting all RNA sequences that have been bound by the tagged RBP. The bound sequences are then separated from the RBP and hybridized against a microarray. As this assay is carried out in a large excess of RNA, the relative abundance of bound sequences indicates relative binding affinity.

**SELEX** [8], or “systematic evolution of ligands by exponential enrichment”, is another in-vitro technique for finding consensus motifs for RBPs. A SELEX run begins with a pool of DNA sequences of equal length, covering all combinatorically possible sequences. The DNA is in-vitro transcribed and an RBP is incubated in the resulting pool of RNA. Once incubation is finished, the unbound RNA sequences are washed out, whereas the bound targets are separated from the RBPs and reverse-transcribed to cDNA, which is then amplified via PCR and serves as the new starting point for the next SELEX run. After a certain number of runs, the resulting “winner” sequences are sequenced.

**High-throughput SELEX** [12] differs from SELEX in that a sample of all bound RNAs is sequenced using high-throughput methods after every run, thus allowing for the estimation of possible alternative binders.

In the context of the complementarity hypothesis and using datasets consisting of RBP-RNA binding motifs, I attempt to investigate theoretical binding affinity of RNA binding proteins towards their own mRNAs. I explore the availability of RNA-binding motifs that are bound by RBPs in every mRNA sequence of the human transcriptome, including the autologous mRNA of each RBP used in the analysis. If RBP motifs are indeed enriched in their own mRNAs, this can point to enrichment in binding affinity between a given RBP and the mRNA coding for it. To determine the binding affinity between an RBP and a given mRNA, the occurrence of RNA-binding motifs within the RNA sequences is investigated using the tool FIMO by MEME Suite [9].

* Short section on results once the final graphs arrive
* Section on how those results are relevant to the scientific community, what to they want to inspire

# Materials and Methods

## Datasets

~~As my analysis is based on the work of Arthur Theuer [11], the natural choice was to continue working with the data he was using. They have led to interesting results with the previous approaches, hence it is to be expected that my analysis, being a generalization of what came before, should point in the same direction. In my work, I focused on using position matrices to determine coverage of a set of RBPs over all sequences in the human, therefore I only used those portions of Arthur Theuer’s datasets that contained matrix data.~~

### RNAcompete [9]

Where did I get it

What’s included (how many proteins, which species, which info on each protein, other info)

How many RBPs for human?

Matrices? How many per protein?

What filtering?

What’s finally included in my analysis?

### SELEX [9]

Where did I get it

What’s included (how many proteins, which species, which info on each protein, other info)

How many RBPs for human?

Matrices? How many per protein?

What filtering?

What’s finally included in my analysis?

### HT-SELEX (2020) [10]

Where did I get it

What’s included (how many proteins, which species, which info on each protein, other info)

How many RBPs for human?

Matrices? How many per protein?

What filtering?

What’s finally included in my analysis?

This study used High-throughput RNA-SELEX to find and validate RNA-binding preferences of RNA binding proteins. It includes a large set of known RBPs and RBP binding domains as well as proteins that have been observed to bind RNA but are not among the canonical class of RBPs. All in all, this study yielded 145 binding models for 86 RBPs. While most of the binding data referred to proteins binding linear RNA sequences, around a third of it gave structural motifs, where RNA was folded when bound.

MOVED FROM INTRO

The ATtRACT-database by Giudice et al. [9] is a unification of several experimentally validated datasets containing RNA binding protein motifs. Being a collection of multiple datasets, ATtRACT includes motifs verified by different experimental techniques. In my analysis, only motifs coming from RNAcompete and SELEX experiments were used, as their respective quality scores are among the highest. All in all, RNAcompete experiments provided 96 different motifs from 77 RBPs, while SELEX contained 46 motifs from 27 proteins.

As my second source, the HT-SELEX [10] experiment by Jolma et al. provided another high-quality collection of protein motif data. The 49 proteins used in my analysis provided 69 different matrices, the length of which varied between three and 21 nucleotides.

As my objective was to find out whether a protein was more likely to bind to its cognate mRNA than to any random mRNA, I used the MANE select database as a background for my analysis. MANE contains the transcripts of around 98 % of all protein-coding genes in the human genome, including the autologous transcripts of all RBPs used in my investigation.

MOVED FROM INTRO

ATtRACT is a project by Guidice et al. aiming to unify a series of different resources for information on RNA binding proteins and experimentally determined binding motifs. For this, the database ATtRACT has been created, which holds information on 370 RNA binding proteins and 1583 RBP motifs from a grand total of 38 different organisms. The motifs vary in length between 4 and 12 nucleotides

The data comes from the databases CISBP-RNA, SpliceAid-F and RBPDB. Not only has a unified database been created, but along with it a search algorithm, allowing for efficient search of motif occurrences in sequences.

* SELEX and RNAcompete have been extracted due to their high-quality scores
* PPMs have been created via alignment of the best-scoring motifs of each experiment;

IUPAC ambiguous letters are treated by counting each possibility as equally likely

* In both RNAcompete and SELEX, a single protein can have multiple motifs:

RNAcompete 77 proteins and 96 motifs

SELEX 49 proteins and 69 motifs

From the ATtRACT database, the RNAcompete and SELEX parts were extracted due to their quality. In RNAcompete, some proteins showed affinity for multiple significantly different motifs. When this was the case, a protein in RNAcompete would yield multiple different matrices with unique matrix IDs. In my analysis, I used all available matrices.

* carried out in siloco analysis of protein-RNA structures available in PDB
* 38 different organisms and length from 4 to 12 nucleotides
* What kind of filtering was done?

**Create plots that show:**

* **lengths of motifs of each experiment**

### MANE select (v0.95) transcriptome

To assess enrichment in autologous binding, a distribution of binding affinities serving as a null hypothesis is necessary. The MANE select database, being a large repository for almost 98 % of the human transcriptome, adeptly fit this purpose. MANE, or “Matched Annotation

from NCBI and EMBL-EBI“, is a large-scale collaboration between the National Center for Biotechnology Information and the European Bioinformatics institute as a Branch of the European Molecular Biology Laboratory. The database is available for download on [the NCBI homepage](https://www.ncbi.nlm.nih.gov/refseq/MANE/).

Its goal was to create a clearly annotated and matching database for human transcriptome data. It now covers over 18.000 transcripts with detailed information on each protein-coding gene’s location, function, 3’-UTR, 5’-UTR and coding sequence (CDS) length. These transcript subsequences serve an important purpose in my analysis, as I consider enrichment of binding in each of them individually. As a filtering step, all RBPs in my dataset also had to have a corresponding autologous mRNA sequence available in MANE, otherwise they were discarded.

## Previous work

Arthur Theuer and Thomas Kapral first initiated the analysis of enrichment in autologous binding of RNA-binding proteins. Their workings constructed a solid foundation for my project and indicate, that there is more information to be extracted from RBP-RNA interactions. While Arthur Theuer’s Bachelor’s thesis explored different frameworks for testing enrichment in autologous binding of RBPs, my project focusses on one of those frameworks, namely the application of position probability matrices to the problem of examining how affine an RBP is to its own mRNA.

### The datasets

The datasets used by my predecessor are largely equal to the ones described in the previous section. However, since I only used RBP data that included probability matrices, the dataset provided by Dominguez et al. had to be excluded from the analysis. This dataset, which contains 76 RBPs and their respective affinity motifs, was created, and results published, in a 2018 study that used High-throughput RNA bind-n-seq to determine RBP motifs. [4]

### Background

As this matching was carried out for every transcript found in the MANE transcriptome database, a coverage value was calculated for each transcript, including the autologous transcript of every RBP.

As compositional and length biases are of concern, the background to be tested against was manipulated in several ways.

Instead of using the transcriptome sequences as they are, in one approach, they were randomly shuffled around to exclude compositional biases. In another approach, only transcripts of equal or higher length than the autologous mRNA were considered and, in turn, randomly shuffled around so as to avoid length and compositional biases. In a third approach, random sequences of equal length to the autologous mRNA of each RBP were created using the di-nucleotide frequency distribution of the MANE database.

### Exact matching

Every RBP in the dataset has at least one motif it shows high affinity for. In an exact matching procedure, an alignment of these “best” motifs to every point along a sequence is attempted and exact matches reported. Using the resulting matches, a coverage value for each RBP-transcript pair can be calculated.

From a probabilistic perspective, motif length plays a big role in an exact matching procedure. Since only an exact match between the protein’s motif and the scanned RNA stretch will contribute to coverage, shorter motifs would be represented considerably more often in any random sequence than longer motifs would. To remedy this bias, motifs were fragmented to a specified size. Whenever a motif was longer than allowed, every possible subsequence of this motif was used to determine a combined coverage value.

### In-silico translation

A different framework for analyzing binding preferences of RBPs is given by a slightly altered research question: “Does an mRNA bind what it codes for?”. Considering this question, the motifs of each RBP were translated into all possible amino acid sequences they might code for. In terms of background, the MANE transcriptome includes the amino acid sequences of each RBP. Again, the motifs, now amino acid sequences, were matched with each protein sequence to determine coverage values.

### Position probability matrices

Probability matrices offer a probabilistic/generalized approach to the match finding problem. In this framework, a matrix is placed against the sequence to test, and a score is calculated by adding up/multiplying the values in the matrix. How exactly this scoring procedure happens is explained in detail in the next section.

### Results

My analysis is based on previous work by Arthur Theuer, who’s investigation is, in turn, derived from Thomas Kapral’s excellent groundwork. Arthur Theuer established several frameworks for investigating this question. In his analysis, he attempted to investigate binding behavior of RBPs by finding exact matches of protein motifs in transcript, he employed in-silico translation and worked with probability matrices to explore his hypothesis.

By using the same datasets as Arthur Theuer, I extended his analysis by an in-depth view of how theoretical protein-RNA binding behavior can be investigated using probability matrices, which offer a probabilistic view on binding affinity.

Chart, diagram

Description automatically generated

Figure 2: The distribution of z-scores over autologous transcripts (left of each section) and the background (right of each section). Exact matching was used to obtain these results. Significant enrichment of autologous binding was observed in most transcript subsequences.

Diagram

Description automatically generated

Figure 3: In-silico translation yielded results showing significant enrichment in autologous binding of most transcript subsequences

## 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:

Table

Description automatically generated

Figure 4: Example for experimental sequences used to construct PFM

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

A picture containing text, clock

Description automatically generated

Figure 5: Position frequency matrix (PFM)

* FIND STUDIES THAT USED PFMs

### Position probability matrix

In a similar fashion 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:

Table

Description automatically generated

Figure 6: Position probability matrix (PPM)

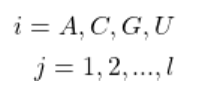
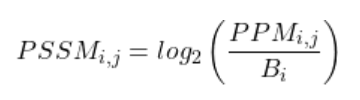
In practice, the position probability matrix can be used to give a probability score to a novel motif by multiplying the letters’ corresponding probabilities over all positions. In a programmatical context, one usually deals with rather long/a large number of motifs to assign a score to. Potential downsides 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:

Figure 7: Constructing a Position-specific scoring matrix by taking the binary logarithm of the likelihood-ratio



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:

A screenshot of a computer

Description automatically generated with low confidence

Figure 8: Position-specific scoring matrix (PSSM)

#### 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 adding “pseudocounts” to the probability values in the PPM. In most practical applications, a pseudocount value of 0.1 is added to each entry in the position probability matrix. Doing this for the above PPM leads to the following PSSM:

Table

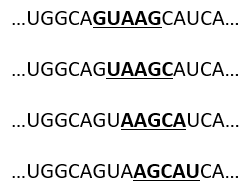
Description automatically generated

Figure 9: Position-specific scoring matrix (PSSM) with pseudocounts

### Scoring an occurrence

Every matrix encodes the affinity of an RBP for a certain motif. While exact motifs can be derived from the matrix, the probabilistic view of affinity is what makes this approach so popular. Instead of looking for an exact overlap between a “best” motif in a sequence, a matrix is “slid” over the sequence and a score is calculated at every position by adding up the value of the respective nucleotide in the matrix. Using the PSSM from above, the following scores can be assigned to the bold-marked sequence positions below:

Figure 10: Scoring procedure for a sequence given a PSSM

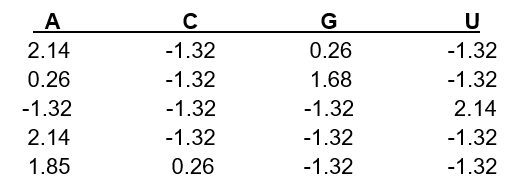


0.26 + (-1.32) + (-1.32) + 2.14 + (-1.32) = **-1.56**

(-1.32) + 0.26 + (-1.32) + (-1,32) + 0.26 = **-3.44**

2.14 + 0.26 + (-1.32) + (-1.32) + 1.85 = **1.61**

2.14 + 1.68 + (-1.32) + 2.14 + (-1.32) = **3.32**



Having calculated the corresponding scores, we cannot really tell whether each score represents a “strong” or a “weak” match of a protein to a motif. To decide the threshold where a score becomes a “good” score, certain statistical methods are necessary.

### Threshold setting

#### P-value approach

* For a given matrix, compute the distribution of all possible scores -> p-value of a given score “s” = 1-cdf(s)
* include distribution of matrix above
* calculate score necessary for a p-value 0.01
* What if matrix is very long?

##### Dynamic programming

* A mathematical approach to solving a certain type of problem
* “stagecoach problem”?
* Consists of splitting a large problem into smaller problems and solving individually

#### Q-value approach

* Benjamini-hochberg
* FDR-rate

#### E-value

* Interpreted as the “corrected” p-value

#### Information content

* in bits
* Stormo and Hertz’ interpretation

### Background distributions

As described in the previous topic, to convert PPMs to PSSMs, the ratio between the likelihood of observing a given letter at a certain position of a matrix and the likelihood of observing a given letter in the background distribution is calculated. The resulting value is the likelihood-ratio, of which then the binary logarithm is computed to receive a log-likelihood ratio for every entry of the matrix. As the background distribution is essential in this calculation, it is necessary to spend some time on the subject.

* Uniform distribution: 0.25 per nucleotide; why not appropriate? di-nucleotide frequency
* Interpretation of log-likelihood in light of background distribution

0 = identical to background

negative = less likely than background would suggest; large negative value would imply that, considering the background, this position should not be as likely as it is.

Is this anything worth considering? If a position scores a large negative value, a two-sided test would yield a significant p-value, even though the protein has an extremely low theoretical affinity

positive = the motif diverges from what the background would suggest; It stands out and may be considered “conserved”, if one only considered the background nucleotide frequency distribution of the scanned sequence.

a high value implies strong theoretical affinity to this position by the protein

Markov chains of zero order

A background can be given via a Markov chain. The zero-order Markov chain only considers the absolute frequencies of each nucleotide

Markov chains of higher order

It is worth thinking about using higher-order Markov chains when constructing a background distribution for motif search. A higher order distribution not only gives the absolute frequencies per nucleotide, but also the probability of a certain nucleotide succeeding another one, e. g. the probability of observing a C after a G.

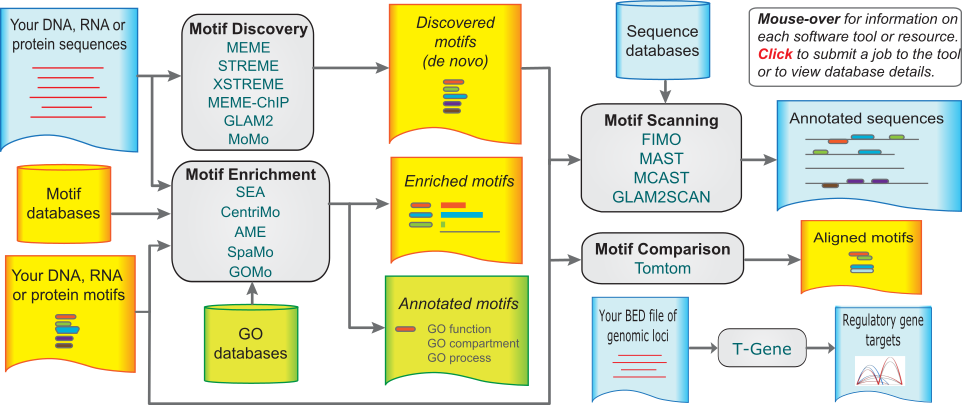
Interesting master’s thesis where someone constructed higher-order PSSMs for higher-order-search

Interesting paper talking about improvements in promoter reg. elements by using higher-order

## FIMO

### The MEME Suite

Meme Suite is a large collaboration comprised of multiple software packages that offer a full pipeline for motif discovery, enrichment, scanning and comparison. The MEME Suite package offers a full pipeline with a web application, a command line tool and an R integration.



### What is FIMO

* Who made FIMO, when was it made, how was it made (approx.)
* What’s the problem it’s trying to solve
* How does it attempt to solve it
* Where has it been used

### FIMO Input preparation

#### Motifs

FIMO takes as an input tab-separated-value files containing position probability matrices with headers containing an identifier, additional information, the matrix length and width and more optional data. At the top of the file, some information on the MEME Suite version used as well as the alphabet the motifs are coding for.

These files are appropriately output by some of the MEME Suites software tools, but can also be created from scratch by following the simple formatting layout provided by FIMO. Alternatively, FIMO offers a variety of conversion tools for common file types encountered in the space of motif analysis.

#### Sequences

The sequences the previously prepared motifs will scan have to be handed to FIMO in “.fasta” or, alternatively, “.fna”-format. Files of this format contain a unique identifier, such as a gene ID, and the sequence in question. In my analysis, I created one such file for each subsequence-type of the transcripts, i. e. 3’-UTR, 5’-UTR and CDS, as well as for the whole transcript itself.

#### Background

**Distribution of nucleotides by subsequence**

A subsequence in this context can be understood as a smaller part of an entire transcript scanned by FIMO, i. e. the 3’-UTR, 5’-UTR and CDS.

As the frequency of nucleotides determines the entries of the PSSM, it has a large impact on the entire analysis.

Chart, pie chart

Description automatically generated

How does the background impact the scores?

Whenever the background frequency is greater than probability of a nucleotide as given by the matrix, the value yielded by the conversion from PPM to PSSM would be a negative one. Therefore, whenever a nucleotide has high frequency in the background, tendentially more values in the PSSM would be negative

### Statistical background

#### Role of the background

In the context of counting motif matches using position probability matrices, an important issue is raised in determining what counts as a “match”. When a position in a sequence is scored according to a certain probability matrix, a wide range of scores can come out. To determine which score represents the lower threshold for what could count as a “match”, a numerical approach was taken in Thomas’ and Arthur’s analyses. In the following paragraph, I will attempt to describe said approach.

A motif of length five would hold 4 x 5 would hold 4^5 possibilities for computable scores. This arises from the fact that each nucleotide can come at any position. For determining a threshold, all possible scores must be computed. By choosing a user-specific percentile-threshold, a large portion of those possible scores can be discarded as insufficient to count as a match. This percentile threshold can, for example, be 0.05. Then, only the top 5 % of all scores will result in a match.

In theory, this approach is statistically sound. In practice, however, creating the entire distribution of scores poses a computational problem as the length of the matrix increases. With a matrix of length 15, 4^15, so more than a billion, scores would have to be computed. To deal with this computational constraint, an approximation of a threshold can be computed. In Arthur’s analysis, all possible scores were computed for motifs of length eight or below. Motifs longer than eight nucleotides long underwent random sampling. This means, that random arrangements of nucleotides of the length of the motif were generated and stored if the number of random arrangements did not surpass the threshold of 4^8. The result of this approximation is that, for each motif, only 4^8 possible scores are considered for creating the distribution.

While the distribution of possible scores for a given motif in place, a cutoff can yet again be chosen and, thereby, the “matching”-threshold determined.

While this is largely sufficient solution to this NP-hard problem, this approach has an obvious downside: As Zhang et al. state in their paper [13], the P-values determined by methods like score approximation can deviate from the real p-value by orders of magnitude. In this example, the reason for this is the following: as soon as the length of the motif surpasses 10, only a sixteenth of all possible scores is considered. At motif length 12, only one in 256 scores is sampled.

However, as Pan and Phan [14] state in their paper, determining p-values by using random sampling rather than computing a full distribution was “verified to be technically identical” to creating the full distribution, as long as sampling size was larger than 4^10.

Use of a background file: a Markov chain is created (zero order) and used to estimate the log likelihood ratio necessary for building up PSSMs from introduced PPMs.

Converts PPMs to PSSMs and runs them over each sequence.

A p-value cutoff is set by the user. In FIMO, the p-value of a match is calculated by creating the score distribution of all possible scores of a motif. The individual score calculated for a certain position on the scanned sequence is converted to a p-value using dynamic programming

#### Calculating p-values

* introduce example for showing calculations
* Graph for score distribution of a given motif
* calculate p-value of a certain score
* describe dynamic programming?
* Q-value

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.

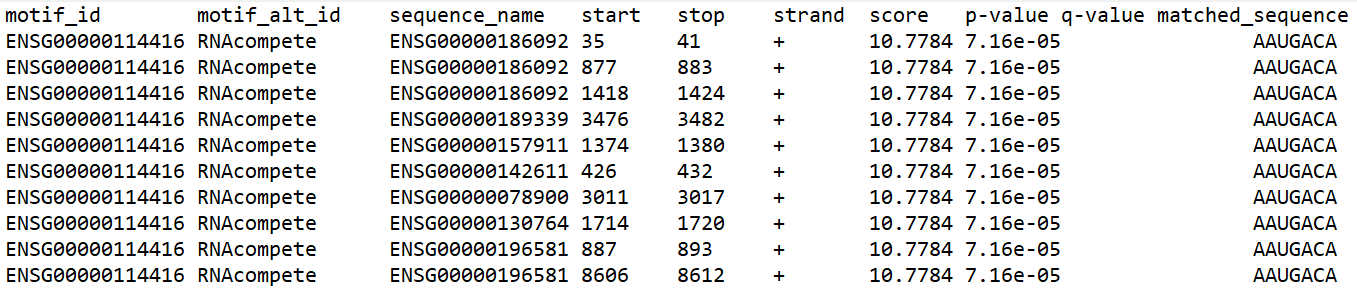
### Outputs

Using the commands as they are elaborated upon in the appendix, one file per analysis is output by FIMO. Each file contains a header and all information on a single significant match in one line. This includes information on identification:

* **the matrix used for scoring**
* **the experiment the matrix originates from**
* **the transcript the match occurred in**

as well as information on the details the match:

* **the position in the full sequence the match started at (start) and ended at (stop)**
* **the strand direction the motif was found in (not used in my analysis)**
* **the exact score achieved by the scored subsequence**
* **the p-value of the score**
* **an empty column for q-values (not computed in my analysis) and**
* **the exact subsequence matched by the motif**



Each output file contains the necessary information for determining, how well any transcript is “covered” by a given matrix. Let’s see how this “coverage” value is determined:

Coverage indicates the theoretical affinity a matrix has towards a given transcript or, in other words, how much of a transcript is “covered” by a given motif. This notion of coverage is generalized by using the probabilistic interpretation of a motif – the position matrix. The exact calculation is best explained by a simple example:

Text

Description automatically generated

Text

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated with medium confidenceText

Description automatically generated

This sequence has a coverage value of 0.45 or, in other words, 45 % of the sequence contain matches to the matrix. In a real-life example, the sequence would be significantly longer, and the coverage value would be expected to be several magnitudes lower.

# Results

The vertical axis of the following plots shows the distribution of z-scores, while the horizontal axis is separated into the three datasets that were used in my analysis. The left side of each part of the triplet visualizes the z-scores of autologous matches in their respective sequence (1. UTR3, 2. CDS, 3. UTR5, 4. full transcript) as well as the mean of these z-scores in the respective color. The right side describes the distribution of z-scores over the entire background, with the mean fixed at zero.

Below each plot, the number of matrices which found matches in the transcriptome is shown contrasted to the total of matrices available in the dataset. The p-values below these numbers stem from a one-tailed z-test describing the probability that the autologous matches’ z-scores came from the same distribution as the background z-scores. As the z-test is one-tailed, the p-value only decreases when the respective autologous z-scores are in a high (i. e. 90%+) percentile of the background distribution.

**P-value cutoff of 1e-4**

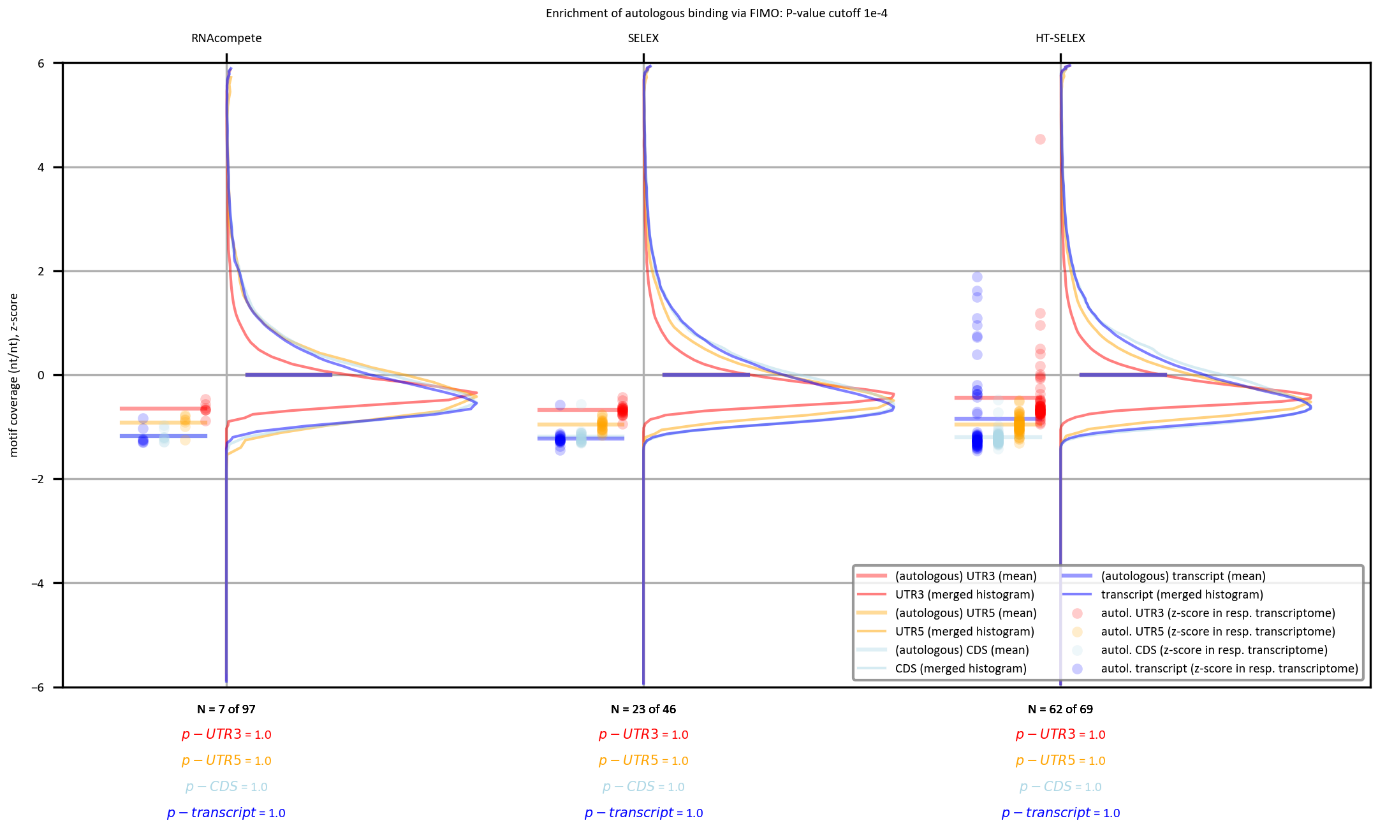


Figure 11: P-value cutoff of 1e-4

The above plotted data originates from a FIMO analysis, in which 96, 46 and 69 matrices from RNAcompete, SELEX and HT-SELEX were used, respectively. The nucleotide frequency distribution provided to FIMO was chosen according to the absolute nucleotide frequencies in the entire transcriptome dataset (MANE select dataset). As a p-value cutoff, FIMO’s default cutoff of 1e-4 was selected.

In RNAcompete and SELEX, all autologous transcripts show lower-than-average coverage in all subsequences, leading largely insignificant p-values. Matrices from HT-SELEX show the same trend, but the coverage of several autologous transcripts exceeded the background-mean, especially in the full-transcript searches.

As given below the figure, the number of matrices that found significant matches in any sequence varies considerably. While in RNAcompete only seven of 97 matrices found matches at all, in SELEX almost half of them did. In HT-SELEX matrices, almost all of them got at least one match.

Chart, diagram

Description automatically generated

Figure 12: P-value cutoff of 1e-3

Chart, diagram

Description automatically generated

Figure 13: P-value cutoff of 1e-2

(Preliminary plots; 5e-2 and 1e-2 analyses that include multiple matrices per protein will follow)

# Discussion

What is obvious from these results is that no enrichment in motif coverage could be observed in autologous transcripts as opposed to the entire transcriptome when the p-value cutoff was too stringent (i. e. 1e-4). This is not surprising, since only a relatively low number of matches was observed in the first place. This is due to fact that a large part of matrices, especially in RNAcompete, had a length covering six nucleotides or less. A matrix of length six can only achieve 4^6 (or 4096) different scores. Since the p-value is calculated via the probability that a certain score occurs in the distribution of all possible scores, this probability being 1/4096, the p-value cutoff of 1/10 000 can never be satisfied. As matrices get longer than seven nucleotides (4^7 = 16 384), surpassing this threshold becomes possible. Remembering the graph that showed the distribution of matrix length per dataset, RNAcompete stayed very much on the shorter side, while SELEX and HT-SELEX proteins recognized longer motifs.

As the p-value threshold became less stringent, more enrichment in autologous binding could be observed.

# References

|  |  |
| --- | --- |
| [1] | F. Crick, "Central Dogma of Molecular Biology," *Nature,* vol. 227, pp. 561-563, 1970. |
| [2] | O. R. J. H. J. M. Michaela Müller-McNicoll, "Auto-regulatory feedback by RNA-binding proteins," *Journal of Molecular Cell Biology,* vol. 11, no. 10, p. 930–939, 2019. |
| [3] | M. C. B. G. W. Y. Meredith Corley, "How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms," *Molecular Cell,* vol. 78, no. 1, pp. 9-29, 2020. |
| [4] | A. G. T. S. A. P. G. R. E. Halbeisen, "Post-transcriptional gene regulation: From genome-wide studies to principles," *Cellular and Molecular Life Sciences,* vol. 65, no. 5, pp. 798-813, 2008. |
| [5] | B. J. Y. J. D. G. Glisovic T, "RNA-binding proteins and post-transcriptional gene regulation," *Journal of the Federation of European Biochemical Societies,* vol. 582, no. 14, pp. 1977-1986, 2008. |
| [6] | A. A. Polyansky and B. Zagrovic, "Evidence of direct complementary interactions between messenger RNAs and their cognate proteins," *Nucleic acids research,* vol. 41, pp. 8434-8443, 2013. |
| [7] | H. K. N. K. Z. H. H. T. M. Q. Ray D, "RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins," *Methods,* pp. 118-119, 2017. |
| [8] | F. S. H. J. N. K. M. S. &. C. S. Yasmeen, "Therapeutic Interventions into Innate Immune Diseases by Means of Aptamers," *Pharmaceutics,* vol. 12, no. 10, p. 955, 2020. |
| [9] | G. G, S.-C. F, T. C and L.-P. E, "ATtRACT-a database of RNA-binding proteins and associated motifs," *Database (Oxford),* 2016. |
| [10] | A. Z. J. M. E. M. E. K. T. L. K. U. Y. Y. Z. F. B. G. M. Q. H. T. R. M. L. J. 3. &. T. J. Jolma, "Binding specificities of human RNA-binding proteins toward structures and linear RNA sequences," *Genome Research,* vol. 30, no. 7, pp. 962-973, 2020. |
| [11] | A. Theuer, "Analysis of autologous binding preferences of RNA-binding proteins," *Bachelor Thesis,* 2022. |
| [12] | C. KB, H. TR and M. QD, "High-throughput characterization of protein-RNA interactions," *Brief Funct Genomics,* vol. 14, no. 1, pp. 74-89, 2015. |
| [13] | J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang and M. Q. Zhang, "Computing exact P-values for DNA motifs," *Bioinformatics,* vol. 23, no. 5, p. 531–537, 2007. |
| [14] | Y. Pan and S. Phan, "Guide to Threshold Selection for Motif Prediction Using Positional Weight Matrix," *Proceedings of the International MultiConference of Engineers and Computer Scientists,* vol. 1, pp. 19-21, 2008. |
| [15] | B. Zagrovic, L. Bartonek and A. A. Polyansky, "RNA-protein interactions in an unstructured context," *FEBS letters,* vol. 592, pp. 2901-2916, 2018. |
| [16] | M. Hlevnjak, A. A. Polyansky and B. Zagrovic, "Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels," *Nucleic acids research,* vol. 40, pp. 8874-8882, 2012. |
| [17] | A. A. Polyansky, M. Hlevnjak and B. Zagrovic, "Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code," *RNA biology,* vol. 10, pp. 1248-1254, 2013. |
| [18] | D. D, F. P, A. MS, S. A, H. M, P. T, B. C, L. NJ, V. N. EL, P. GA, Y. GW, G. BR and B. CB, "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins," *Mol Cell,* vol. 70, no. 5, pp. 854-867, 2018. |
| [19] | A. C. a. F. H.-T. Allain, "From Structure to Function of RNA Binding Domains," in *Madame Curie Bioscience Database [Internet]*, Austin (TX), Landes Bioscience, 2000-2013. |
| [20] | R. M. V. J. D. F.-K. P Andrew Chong, "RGG/RG Motif Regions in RNA Binding and Phase Separation," *Journal of Molecular Biology,* vol. 430, no. 23, pp. 4650-5685, 2018. |

# Acknowledgements

# Appendix

**Setting path for FIMO executable in command line**

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

To save space, the “—text” option in the FIMO command enables us to write the tab-separated data directly to the appropriate file.

First, make sure to create a .tsv file for each FIMO command you will run, e. g. for each output file you expect to have. In my case, this would be 12 files, each called “experiment\_subsequence.tsv”. The commands to create these files are best executed from the “DATA” folder in my hierarchy, which is the main branching point for all the data FIMO needs. You can use the following commands to create the files:

cd DATA/FIMO\_OUT

mkdir pval5e-2

touch pval5e-2/rnacompete\_UTR3.tsv

Once the directories and files are in place, the FIMO commands are used as such:

fimo

--text

--max-stored-scores 2,147,483,646

--thresh 5e-2

--bfile FIMO\_input/background\_file.txt

FIMO\_input/motifs/motif\_file.txt

FIMO\_input/sequences/transcript\_file.txt

> FIMO\_OUT/ target\_file.tsv

Let’s break down each part of this command:

--text: reduces the output to “tsv” output sent to the standard-output, i. e. to the command line itself. This argument allows us to redirect the outputs to the appropriate file using the “>” operator.

--max-stored-scores: for score significance to be calculated, a certain number of matches will have to be held in memory. This argument is especially important when working with Q-values instead of P-values. The large number after the command is the maximum number of stored scores FIMO allows for.

--thresh: sets the custom P-value cutoff. The default cutoff lies at 1e-4.

--bfile: after this argument, a file containing the nucleotide frequency distribution of the background is added – in my example called “background\_file.txt”. This file can be created using the “fasta-get-markov” command, which is distributed together with MEME Suite. More on this command below. The background file contains a zero-order Markov model of the background distribution. No higher order models are allowed as input to FIMO. If a higher order model is entered, FIMO will only consider the zero-order part.

The penultimate argument hands the input motif file to FIMO, while the last one enters the sequence file.

The “>”-operator: In Linux, this operator is often used to write the output of a command to a file. Additionally, this operator overwrites the contents of the file, if they exist.

For each experiment-subsequence combination (e. g. RNAcompete+CDS, SELEX+UTR3 etc.) such a command must be constructed. The choice of background file and P-value threshold is left up to the user. To chain up multiple of these commands, one can place a semicolon “;” or an “and”-operator “&&” in between. While for the “and”-operator, the preceding command must have been completed successfully, the semicolon chains up commands without requiring successful completion.

Here are all FIMO commands I used, chained up by semicolons:

cd ~; cd Moritz\_BSc/BSc\_enriched\_autologous\_RBP/DATA/

Creation of file containing background nucleotide frequencies:

fasta-get-markov sequences/fimo\_transcriptome\_full.txt

folder system by cutoff:

pval5e-2

pval1e-2

pval1e-3

pval1e-4

cd FIMO\_OUT

mkdir pval5e-2

touch pval5e-2/rnacompete\_UTR3.tsv

touch pval5e-2/rnacompete\_CDS.tsv

touch pval5e-2/rnacompete\_UTR5.tsv

touch pval5e-2/rnacompete\_full.tsv

touch pval5e-2/selex\_UTR3.tsv

touch pval5e-2/selex\_CDS.tsv

touch pval5e-2/selex\_UTR5.tsv

touch pval5e-2/selex\_full.tsv

touch pval5e-2/htselex\_UTR3.tsv

touch pval5e-2/htselex\_CDS.tsv

touch pval5e-2/htselex\_UTR5.tsv

touch pval5e-2/htselex\_full.tsv

mkdir pval1e-2

touch pval1e-2/rnacompete\_UTR3.tsv

touch pval1e-2/rnacompete\_CDS.tsv

touch pval1e-2/rnacompete\_UTR5.tsv

touch pval1e-2/rnacompete\_full.tsv

touch pval1e-2/selex\_UTR3.tsv

touch pval1e-2/selex\_CDS.tsv

touch pval1e-2/selex\_UTR5.tsv

touch pval1e-2/selex\_full.tsv

touch pval1e-2/htselex\_UTR3.tsv

touch pval1e-2/htselex\_CDS.tsv

touch pval1e-2/htselex\_UTR5.tsv

touch pval1e-2/htselex\_full.tsv

mkdir pval1e-3

touch pval1e-3/rnacompete\_UTR3.tsv

touch pval1e-3/rnacompete\_CDS.tsv

touch pval1e-3/rnacompete\_UTR5.tsv

touch pval1e-3/rnacompete\_full.tsv

touch pval1e-3/selex\_UTR3.tsv

touch pval1e-3/selex\_CDS.tsv

touch pval1e-3/selex\_UTR5.tsv

touch pval1e-3/selex\_full.tsv

touch pval1e-3/htselex\_UTR3.tsv

touch pval1e-3/htselex\_CDS.tsv

touch pval1e-3/htselex\_UTR5.tsv

touch pval1e-3/htselex\_full.tsv

mkdir pval1e-4

touch pval1e-4/rnacompete\_UTR3.tsv

touch pval1e-4/rnacompete\_CDS.tsv

touch pval1e-4/rnacompete\_UTR5.tsv

touch pval1e-4/rnacompete\_full.tsv

touch pval1e-4/selex\_UTR3.tsv

touch pval1e-4/selex\_CDS.tsv

touch pval1e-4/selex\_UTR5.tsv

touch pval1e-4/selex\_full.tsv

touch pval1e-4/htselex\_UTR3.tsv

touch pval1e-4/htselex\_CDS.tsv

touch pval1e-4/htselex\_UTR5.tsv

touch pval1e-4/htselex\_full.tsv

**Pvalue: 1e-4**

**ATtRACT - SELEX**

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt > FIMO\_OUT/pval1e-4/selex\_UTR5.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt > FIMO\_OUT/pval1e-4/selex\_CDS.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt > FIMO\_OUT/pval1e-4/selex\_UTR3.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_selex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt > FIMO\_OUT/pval1e-4/selex\_full.tsv

**ATtRACT - RNAcompete**

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt > FIMO\_OUT/pval1e-4/rnacompete\_UTR5.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt > FIMO\_OUT/pval1e-4/rnacompete\_CDS.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt > FIMO\_OUT/pval1e-4/rnacompete\_UTR3.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_attract\_rnacomp.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt > FIMO\_OUT/pval1e-4/rnacompete\_full.tsv

**HT-SELEX**

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_5utr.txt > FIMO\_OUT/pval1e-4/htselex\_UTR5.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_cds.txt > FIMO\_OUT/pval1e-4/htselex\_CDS.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_3utr.txt > FIMO\_OUT/pval1e-4/htselex\_UTR3.tsv;

fimo --text --max-stored-scores 21474836466 --thresh 1e-4 --bfile FIMO\_input/markov\_full.txt FIMO\_input/motifs/fimo\_htselex.txt FIMO\_input/sequences/fimo\_transcriptome\_full.txt > FIMO\_OUT/pval1e-4/htselex\_full.tsv

**Pvalue: 1e-3**

**ATtRACT - SELEX**

fimo --text --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\_OUT/pval1e-3/selex\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-3/selex\_CDS.tsv;

fimo --text --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\_OUT/pval1e-3/selex\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-3/selex\_full.tsv

**ATtRACT - RNAcompete**

fimo --text --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\_OUT/pval1e-3/rnacompete\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-3/rnacompete\_CDS.tsv;

fimo --text --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\_OUT/pval1e-3/rnacompete\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-3/rnacompete\_full.tsv

**HT-SELEX**

fimo --text --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\_OUT/pval1e-3/htselex\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-3/htselex\_CDS.tsv;

fimo --text --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 > FIMO\_OUT/pval1e-3/htselex\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-3/htselex\_full.tsv

**Pvalue: 1e-2**

**ATtRACT - SELEX**

fimo --text --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\_OUT/pval1e-2/selex\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-2/selex\_CDS.tsv;

fimo --text --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\_OUT/pval1e-2/selex\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-2/selex\_full.tsv

**ATtRACT - RNAcompete**

fimo --text --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\_OUT/pval1e-2/rnacompete\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-2/rnacompete\_CDS.tsv;

fimo --text --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\_OUT/pval1e-2/rnacompete\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-2/rnacompete\_full.tsv;

**HT-SELEX**

fimo --text --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\_OUT/pval1e-2/htselex\_UTR5.tsv;

fimo --text --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\_OUT/pval1e-2/htselex\_CDS.tsv;

fimo --text --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 > FIMO\_OUT/pval1e-2/htselex\_UTR3.tsv;

fimo --text --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\_OUT/pval1e-2/htselex\_full.tsv

**Pvalue 5e-2**

**ATtRACT - SELEX**

fimo --text --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\_OUT/pval5e-2/selex\_UTR5.tsv;

fimo --text --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\_OUT/pval5e-2/selex\_CDS.tsv;

fimo --text --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\_OUT/pval5e-2/selex\_UTR3.tsv;

fimo --text --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\_OUT/pval5e-2/selex\_full.tsv

**ATtRACT - RNAcompete**

fimo --text --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\_OUT/pval5e-2/rnacompete\_UTR5.tsv;

fimo --text --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\_OUT/pval5e-2/rnacompete\_CDS.tsv;

fimo --text --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\_OUT/pval5e-2/rnacompete\_UTR3.tsv;

fimo --text --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\_OUT/pval5e-2/rnacompete\_full.tsv

**HT-SELEX**

fimo --text --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\_OUT/pval5e-2/htselex\_UTR5.tsv;

fimo --text --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\_OUT/pval5e-2/htselex\_CDS.tsv;

fimo --text --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 > FIMO\_OUT/pval5e-2/htselex\_UTR3.tsv;

fimo --text --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\_OUT/pval5e-2/htselex\_full.tsv

**EXTRA WRITTEN STUFF**

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

**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.