BACHELOR THESIS

RNA binding motifs reveal tendencies towards autogenous binding

submitted by
Moritz Haderer

in partial fulfillment of the requirements for the degree of
Bachelor of Science (BSc)

Vienna, 2022

Degree programme code as it appears on the student record sheet:	**UA 033 630**
Degree programme as it appears on the student record sheet:	**Bachelor's degree programme Biology**
Supervisor:	**Univ.-Prof. Dr. Bojan Zagrovic, BA**

Abstract

The central dogma of molecular biology [1] is an accepted model for the passing down of genetic information in the cell. While it captures the essence of life on a molecular level, there are several types of interactions that remain incompletely understood when viewed from this perspective. Interactions between RNA-binding proteins (RBPs) and RNA count among these. They are integral parts of many cellular processes such as post-transcriptional regulation, RNA transport- and localization as well as post-transcriptional modification of RNA. As RBPs play a guiding role in these interactions, their availability needs to be tightly regulated so as to ensure proper cellular functioning. Autoregulatory feedback between RBPs and their own mRNAs [2] often lies at the heart of these regulatory loops.

The autogenous mRNA/protein-complementarity hypothesis, suggesting that proteins in general bind in a complementary, co-aligned manner to their own mRNAs, helps to understand how such interactions between proteins and their own mRNA can occur. As it was only recently proposed, however, the complementarity hypothesis suffers from a lack of extensive experimental results. The present work leverages recent experimental data on the preferred RNA motifs of different RBPs to provide a systematic test of the complementarity hypothesis.

The results of the analysis show a general enrichment of preferred RNA motifs in the autogenous mRNAs of different RBPs. The agreement is, in particular, greater when an RBP's binding preferences are relatively unspecific. While protein-RNA interactions can be highly specific, these more vague interaction patterns point towards the matching of certain more coarse-grained physicochemical properties of the interacting 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 autogenous mRNA in a global manner and, therefore, bind and regulate it.

Contents

1. Introduction

RNA binding proteins (RBPs) are a class of proteins that interact directly with RNA molecules [3]. These interactions have been found to be involved in different 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 of great importance for the proper functioning of cells, our understanding of the interactions between these key proteins remains incomplete.

The recently proposed autogenous mRNA/protein-complementarity hypothesis creates a framework for understanding these protein-RNA interactions in a broader sense. According to the hypothesis, proteins and their autogenous mRNA are physicochemically complementary to each other and bind in a co-aligned manner, especially if unstructured [6] [7]. The hypothesis is a generalization of the known stereochemical hypothesis of the origin of the genetic code, which states that codon assignments evolved from direct binding preferences between amino acids and codons [8] [9]. Moreover, whenever genetic information flows from mRNA to protein, the physicochemical properties of the mRNA sequence are translated to the physicochemical properties of the protein. In the context of the complementarity hypothesis, a stretch of mRNA that codes for a given RNA-binding domain within a protein would transfer a certain affinity towards itself via the process of translation. In this context, it has been found that translation does not just consist of going from a sequence of bases a sequence of amino acids, but involves passing down physicochemical properties like hydrophobicity profiles and affinity for certain nucleobases [7]. This phenomenon could provide a foundation for understanding the affinity of proteins for their cognate mRNAs and beyond [10].

1.1. RNA-binding domains

During evolution, RBPs got more and more specialized to recognize specific RNAs [5], developing binding regions that show affinity only to specific RNA recognition patterns.

These binding regions, also called binding domains, have developed affinity for relatively precise sets 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 unfolded RNA molecules, while 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**. [11]

The RNA-Recognition Motif is found in 0.5%-1% of human genes and consists of a stretch of around 90 amino acids [11]. 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.

The Double-stranded RNA-binding motif prefers to bind double-stranded RNAs [11]. This also implies that these interactions are less specific, as the base-information is not as readily available as in a single-stranded RNA 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 [11]. 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 Arginine and Glycine residues [12]. 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.

1.2. Data acquisition methods

As experimental support for the complementarity hypothesis is scarce, we resort to in-silico analysis of protein-RNA interaction data for insight. The emergence of high-precision technologies like RNAcompete and SELEX, the methods that enable the determination of the exact nucleotide motifs a given RBP can recognize, facilitates this type of analysis significantly.

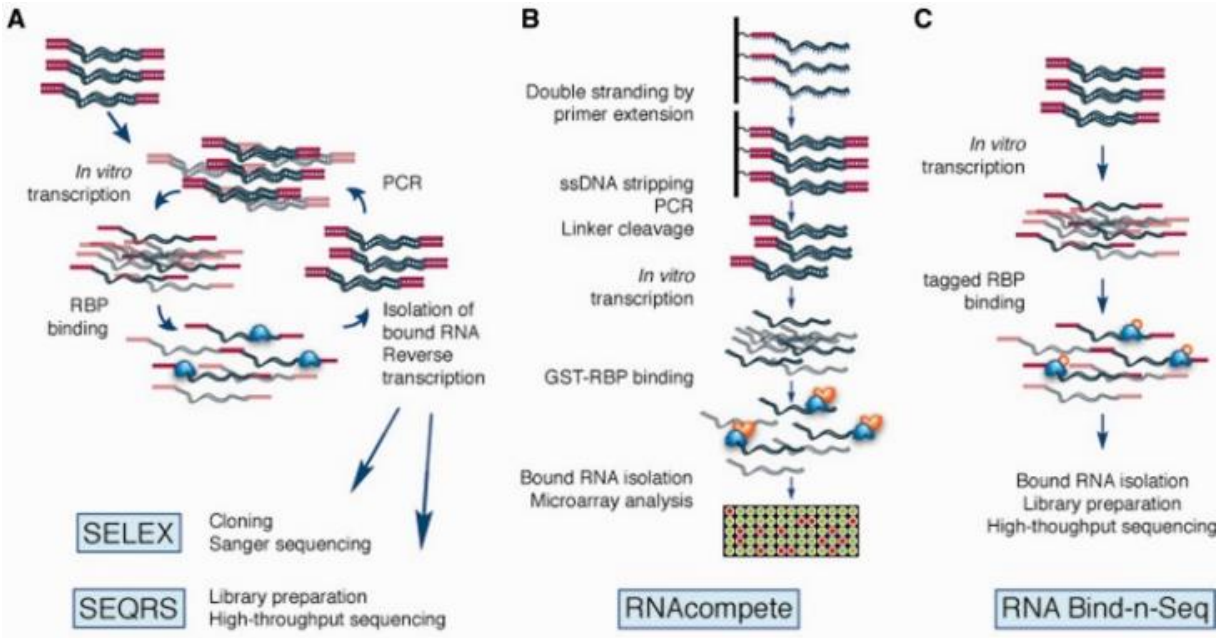

Figure 1: *Description of experimental techniques for RBP binding preference determination* ***[13]***

RNAcompete [14] 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 combinatorially possible sequences of up to 9 bases. After incubation of the RBP in an environment with an excess of RNA, a pull-down assay is performed, capturing 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 [15], or "systematic evolution of ligands by exponential enrichment", is an 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 [13] 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.

1.3. General workflow

Using datasets consisting of RBP-RNA binding motifs, an analysis of the enrichment of RBP binding sites in autogenous RNAs is attempted in the context of the complementarity hypothesis. Position probability matrices (PPMs) and the closely related position specific-scoring matrices (PSSMs) are used to provide a measure of enrichment of motifs. Probability matrices offer a probabilistic approach to the sequence motif match finding problem. Matrices are created from experimentally determined motifs. As RBPs can bind motifs with a certain degree of variability, the probability of a certain nucleotide appearing at any position can be calculated and combined to form a position probability matrix. When a matrix is placed against the sequence to test, a score is calculated by adding up/multiplying the values in the matrix that correspond to the sequence of letters. How exactly this scoring procedure happens is explained in detail in the extended introduction.

The presence of experimentally validated RNA-binding motifs within sequences of the human transcriptome is interpreted as an indication of binding propensity of a given protein towards a transcript. As the set of tested transcripts includes all autogenous mRNAs of the proteins under investigation, a statistical test can reveal an enrichment such propensity. If the known RBP motifs are indeed enriched in their own mRNAs, this may further support the findings that RBPs play a role in the regulation of their autogenous transcripts. The frequency of occurrence of RNA-binding motifs within the RNA sequences was determined using the tool FIMO by MEME Suite [16].

1.4. Results

Three different approaches were taken to study the frequency of occurrence of the known binding motifs of RBPs in question in their autogenous mRNAs. First, scoring matrices were employed via FIMO to calculate a coverage value for each protein in the dataset. Multiple scoring matrices per protein were used to capture, in a probabilistic manner, binding motifs when available and the resulting coverages were normalized by the number of matrices. Second, the same approach was taken only using a single matrix per protein in order to eliminate possible biases created by non-homogeneous utilization of matrices. Third, an exact-matching procedure was applied to the same dataset. The investigation revealed significant enrichment of known binding motifs within the 3'-UTR and the full transcript of autogenous mRNAs of RBPs in the first two approaches, while enrichment of nearly all subsequences could be observed in the exact matching analysis.

The robustness of these results, considering the probabilistic and generalized methods utilized to test the effect, supports the hypothesis that proteins may show affinity towards their autogenous mRNAs. However, a further in-depth investigation is necessary to reveal exactly how and when these important interactions occur and how they affect cellular processes.

2. Materials and Methods

2.1. Datasets

2.1.1. RNAcompete [17]

The ATtRACT database by Giudice et al. (downloaded at ATtRACT-database) contains RNAcompete results for 194 RBPs from 24 different organisms. For each RBP, at least one consensus sequence is provided as well as information on the protein domain involved in binding. RBPs are labelled by their gene-IDs according to Ensembl [18], Xenbase [19]or the European Nucleotide Archive [20]and gene names as provided in UniProt [21].

The data was filtered to include only human RBP motifs. As proteins can bind different motifs with varying consistency, the database includes a quality score for each motif. This quality score has been calculated based on the probability of observing a given motif in an experiment and can be used as a numerical measure of binding propensity ranging between 0 and 1**. Only the motifs with quality scores of 1** were included. The resulting entries each contained a matrix ID referencing a position-probability matrix, yielding a total of 96 matrices for 77 proteins.

2.1.2. SELEX [17]

The ATtRACT database also contains SELEX results for 41 RBPs from 9 different organisms. The sequences obtained from a SELEX run can be ranked according to their preference to bind RBPs. When variability between sequences is too high, multiple PPMs are calculated for a protein in order to accurately describe the protein's binding preferences. ATtRACT does not provide information on the measure of sequence variability used in matrix construction.

As for RNAcompete, SELEX data was filtered to only include matrices coming from experiments using human RBPs. Furthermore, matrices that are not based on the highest-rated binders are filtered out. After filtering, 26 RBPs remained, many of which bound to multiple significantly different motifs. A total of 46 PPMs could be used in the analysis.

2.1.1. HT-SELEX (2020) [22]

Jolma et al. carried out a large high-throughput RNA-SELEX study with the goal of discovering and validating RNA binding motifs of human RNA binding proteins. The resulting data can be downloaded from the European Nucleotide Archive. It includes known RBPs, isolated 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 proteins. The database contains information on the structural preferences of specific proteins, indicating whether they bound to linear RNA rather than folded RNA. Furthermore, some RBPs have been found to bind RNA as dimers.

As the goal of this analysis is to explore motif coverage over linear RNA sequences of single RBPs, all dimeric binders and RBPs with structural preferences were excluded from the dataset. Additionally, one protein was not available in the MANE select database (as

described below) For each protein, at least one position probability matrix (PPM) was provided, in the end resulting in 49 proteins yielding 69 matrices.

2.1.2. MANE select (v0.95) transcriptome

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.

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. The MANE database represents the transcriptome with equal weight per gene, rather than weighting by the number of transcript variants. Due to this weighting, only the best supported transcript for each gene is included.

2.2. Methodology

2.2.1. Motif occurrence null model

To assess enrichment in autogenous binding propensity, a distribution of the background frequencies of the occurrence of binding motifs, 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. As this analysis explores the frequency of preferred binding motifs of RBPs in their autogenous mRNAs, the equivalent frequencies in mRNAs in general must be established. With a null-model in place, p-values for autogenous interactions can be derived.

2.2.2. Sequence motif search

In a sequence motif search procedure, an alignment of RNA-binding motifs to every subsequence along a transcript is attempted and exact matches reported. From a probabilistic perspective, motif length plays a big role in an exact matching procedure. Since only the exact alignment of a motif to the scanned RNA stretch counts as a match, 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 exceeded the desired length, every possible subsequence of this motif was used to determine a combined coverage value. The matches of individual motif fragments were finally combined, showing matches per complete motif.

2.2.3. Sequence motif search – FIMO [16]

FIMO performs sequence motif search over a set of sequences using PPMs provided. Every matrix is run over every sequence, yielding a score for every position the matrix is compared against as well as a p-value corresponding to the score. A match is reported whenever the calculated p-value is lower than the p-value cutoff specified by the user. For short motifs, a stringent p-value cutoff can lead to incomplete results. This is due to the fact that matrices with a length of 6 nucleotides or less 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 $\frac{1}{4096}$, the p-value cutoff of, for example, $\frac{1}{10\,000}$ can never be satisfied. As matrices get longer than 7 nucleotides (4^7 = 16 384), surpassing this threshold becomes possible.

The motif data coming from RNAcompete, SELEX and HT-SELEX experiments as well as the MANE select database were written to text files according to the format specifications FIMO uses and motif search was performed for a variety of p-value cutoffs. As FIMO uses position-specific scoring matrices (PSSMs) rather than PPMs (a detailed explanation of matrix types can be found in the Extended introduction), a nucleotide background distribution of the scanned transcripts should be provided. This background distribution was chosen to be a zero-order Markov model based on the entire MANE select transcriptome data. Finally, FIMO returns a list of matches, each entry of which contains all necessary identifiers, the p-value of the score and the start and the stop indices of the match in the transcript

2.2.4. Evaluation of motif-occurrence enrichment

Sequence motif search was carried out for every transcript found in the MANE transcriptome database. Any significant motif-sequence alignments were reported as matches, which were

then combined to a coverage value. Coverage indicates how often a given motif is present in a transcript sequence and is calculated as portrayed by the following example:

Sequence of length 40 containing three matches of length six:

AG**UUCAGU**G**UGCAGA**AACUUCGCUUGAAGC**CCCAGU**GCGU

↓

Sequence is converted into an array of zeros of length 40:

00**000000**0**000000**0000000000000000**000000**0000

↓

Matched positions are converted to ones:

00**111111**0**111111**0000000000000000**111111**0000

↓

$$3_{\text{matches}} \cdot 6_{\text{nt}} = 18 \qquad \frac{18}{40} = 0.45$$

The given 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.

Enrichment in motif-occurrence is evaluated by comparing the coverage value of each transcript to the mean coverage a motif achieves. To this end, a z-score for every transcript is calculated.

$$z = \frac{cov_{transcript} - cov_{\mu}}{cov_{\sigma}}$$

Equation 1: *The z-score is determined using this equation. $cov_{transcript}$ is the coverage value of a given transcript, cov_{μ} is the mean coverage over all transcripts and cov_{σ} is the standard deviation over all coverage values.*

2.3. Extended introduction

2.3.1. Position frequency matrices

Different experimentally derived motifs that are bound by a given RBP are often similar in length and composition. To capture the variability between bound motifs, the Position Frequency Matrix (PFM) visualizes the number of times a given nucleotide is 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...

Figure 2: *Example for experimental sequences used to construct PFM*

Counting the number of times nucleotides are observed at a given position, a PMF of the following structure can be constructed:

A	C	G	U
10	0	0	0
2	1	7	0
0	0	0	10
10	0	0	0
8	2	0	0

Figure 3: *Position frequency matrix (PFM)*

2.3.2. Position probability matrices

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 are created by dividing each frequency count by 10:

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

Figure 4: *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 corresponding probabilities over all positions. Computationally, one usually deals with rather long motifs or a large number of motifs to assign a score to. Potential downsides of this type of matrices are 1) multiplication is a more expensive computation than, e. g., addition. Once the sequences to be scored exceed a certain length, this leads to a noticeable increase in the computational burdens. Another downside is that multiplication of very long matrices can lead to underflow when computing the score of a motif. When multiplying many small numbers, floating point precision can be thought of as a resource. This effect might skew the scoring and, consequently, determination of the sufficiency of a score.

2.3.3. Position-specific scoring matrices

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 would suggest, thus reducing its weight.

$$PSSM_{i,j} = log_2\left(\frac{PPM_{i,j}}{B_i}\right) \qquad B = [0.25, 0.25, 0.25, 0.25]$$
$$i = A, C, G, U$$
$$j = 1, 2, ..., l$$

Figure 5: *Constructing a position-specific scoring matrix by taking the binary logarithm of the likelihood-ratio. In this example, a uniform background distribution of 0.25 per nucleotide is assumed. 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	C	G	U
-1.32	-inf	-inf	-inf
-0.32	-1.32	1.49	-inf
-inf	-inf	-inf	2
2	-inf	-inf	-inf
1.68	-0.32	-inf	-inf

Figure 6: *Position-specific scoring matrix (PSSM)*

2.3.3.1. 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:

A	C	G	U
2.14	-1.32	0.26	-1.32
0.26	-1.32	1.68	-1.32
-1.32	-1.32	-1.32	2.14
2.14	-1.32	-1.32	-1.32
1.85	0.26	-1.32	-1.32

Figure 7: *Position-specific scoring matrix (PSSM) with pseudocounts*

2.3.4. 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 binding propensity 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:

A	C	G	U
2.14	-1.32	0.26	-1.32
0.26	-1.32	1.68	-1.32
-1.32	-1.32	-1.32	2.14
2.14	-1.32	-1.32	-1.32
1.85	0.26	-1.32	-1.32

…UGGCA**GUAAG**CAUCA…	0.26 + (-1.32) + (-1.32) + 2.14 + (-1.32) = **-1.56**
…UGGCAG**UAAGC**AUCA…	(-1.32) + 0.26 + (-1.32) + (-1,32) + 0.26 = **-3.44**
…UGGCAGU**AAGCA**UCA…	2.14 + 0.26 + (-1.32) + (-1.32) + 1.85 = **1.61**
…UGGCAGUA**AGCAU**CA…	2.14 + 1.68 + (-1.32) + 2.14 + (-1.32) = **3.32**

Figure 8: *Scoring procedure for a sequence given a PSSM*

Having calculated the corresponding scores, it is hard to say whether each score represents a "strong" or a "weak" match of a protein to a motif. To set the threshold where a score is deemed sufficient, certain statistical methods are necessary.

2.3.5. Threshold setting

A motif of length five contains 4^5 computable scores. This arises from the fact that each nucleotide can come at any position. By computing all possible scores a matrix can yield, a score distribution is created. The PSSM from the example above leads to the distribution in Figure 9. When a percentile significance threshold is chosen, e. g. 0.05, then all scores that lie within the top percentage of e. g. 5 % in this example, of the distribution will be considered sufficient.

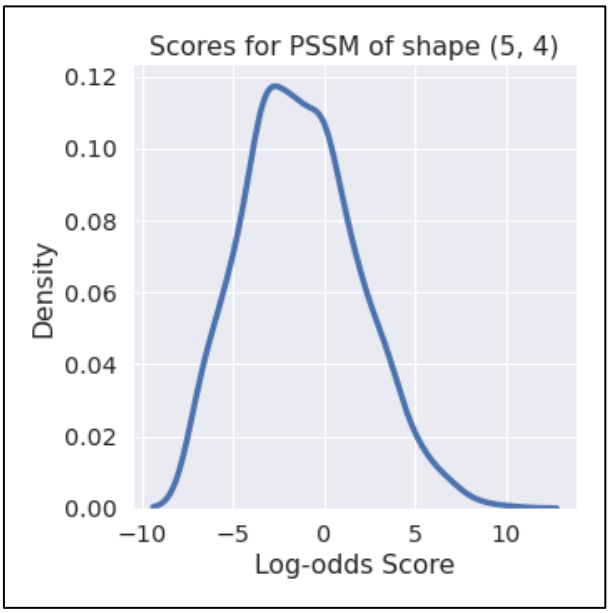

Figure 9: *Score distribution of PSSM.*

While this approach is statistically sound, 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}, i. e. more than 1 billion, scores would have to be computed. To deal with this computational constraint, an approximation of the matrix-specific threshold can be computed by randomly sampling the possible nucleotide arrangements. As Pan and Phan [23] state in their paper, determining score thresholds by using random sampling rather than computing a full distribution was "verified to be technically identical" to creating the full distribution, if the space to be sampled was larger than 4^{10}

2.4. FIMO [16]

FIMO is a tool offering discovery of motif occurrences in sequences as a part of the MEME Suite package [24]. It is used for scanning of sequences with a set of motif matrices and returns all significant matches that may occur. A significance threshold can be set by the user.

2.4.1. FIMO Input

2.4.1.1. 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 is given.

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.

2.4.1.2. Sequences

The sequences must 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.

2.4.1.3. Background nucleotide frequency distribution

When PSSMs are used for discovery of motif occurrences, the distribution of nucleotide frequencies in the searched sequence plays a vital role. As discussed in the previous section, these nucleotide frequencies are involved in computing the PSSM entries from a PPM, yielding a positive value when the probability of a given nucleotide in the PPM is lower than the background probability of the same nucleotide and a negative value when the opposite is the case. An interpretation of this effect is that positive values in a PSSM indicate that the motif diverges from the set of motifs the background distribution would suggest are most likely.

2.4.2. FIMO output

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
- the exact score achieved by the scored subsequence
- the p-value of the score
- an empty column for q-values (not computed in this example) and
- the exact subsequence matched by the motif

motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence
ENSG00000114416	RNAcompete	ENSG00000186092	35	41	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000186092	877	883	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000186092	1418	1424	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000189339	3476	3482	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000157911	1374	1380	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000142611	426	432	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000078900	3011	3017	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000130764	1714	1720	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000196581	887	893	+	10.7784	7.16e-05		AAUGACA
ENSG00000114416	RNAcompete	ENSG00000196581	8606	8612	+	10.7784	7.16e-05		AAUGACA

Figure 10: *The output of a FIMO analysis. The contents of this table are described in detail above the figure.*

3. Results

3.1. Enriched occurrence of motif sequences in RBP's autogenous transcripts

The hypothesis that proteins have affinity towards their own coding sequences, the complementarity hypothesis emphasizes the importance of autogenous feedback in post-transcriptional regulation of RBP-coding genes.

To test this hypothesis, the enrichment of RBP affinity motifs in autogenous mRNA sequences was analyzed. The ATtRACT database [17] as well as the HT-SELEX experiment of Jolma et al. [22] are some of the most comprehensive sources for RBP motif data available, while the MANE select database serves as a null-model for motif enrichment in the entire transcriptome. FIMO [16] was used to scan the transcriptome sequences for motif matches. Once motif-sequence matches were found, the density of motif occurrences in each transcript sequence was calculated and autogenous enrichment was determined using a z-test. Depending on the p-value cutoff specified beforehand, stringency could be regulated. A lower cutoff (e. g. of 1e-4) includes sequence-motif matches that can be interpreted as highly specific, whereas a less stringent one allows for more fuzzy interactions to be considered significant. As different levels of stringency were of interest, each analysis was carried out at a p-value cutoff of 1e-4, 1e-3, 1e-2 and 0.05.

3.1.1. Normalization by the number of matrices per protein

The availability of multiple matrices for certain RBPs posed a difficult question: How can a maximum of binding information per protein be analyzed while avoiding skewing the distribution of z-scores? A possible solution involved considering the RNA-binding domain the motifs were bound by. This approach shifts the focus from the RBP itself to the domains that were identified to be involved in sequence binding during experiments. By putting binding domains at the center of the analysis, an additional layer of information can be extracted from the data, namely the incidence of motifs preferred by a given binding domain, rather than the entire RBP. If multiple matrices are available for a protein, the coverage of the binding

domains in each transcript could be normalized by the number of matrices used. In ATtRACT, different binding domains are stated for some proteins, while the HT-SELEX dataset contained no such information, therefore making a normalization by binding domain impossible.
As an alternative solution to eliminating the biases created by certain proteins containing multiple binding motifs, the calculated coverages for each matrix were combined and divided by the number of matrices.

Using this method at a p-value cutoff of 0.05 (Figure 15), significant increase of binding motif coverage in autogenous mRNAs could be observed in two of three experiments. RNAcompete and HT-SELEX motifs were found considerably more often in autogenous mRNAs as opposed to the entire transcriptome. In particular, the 3'-UTR and the coding sequence (CDS) show a significantly higher incidence of their cognate proteins' binding motifs. While SELEX shows no significant enrichment (p-value above 0.05), the RBP motifs exhibited a higher incidence in their autogenous CDS regions than in the entire transcript. This effect might be explained, firstly, by the nature of PSSMs and, secondly, by the way FIMO measures significance. A PSSM's entries are computed taking into the account the distribution of nucleotide frequencies in the background. All analyses carried out are based on PSSMs created using the nucleotide frequency distribution of the full transcriptome. The results show no significant enrichment in motif coverage in the full transcript sequences, indicating that the motifs investigated do not at high enough density, considering the length of the transcript. As neither the 3'-UTR nor the 5'-UTR show enrichment, being shorter subsequences, the results cannot stem from a uniformly distributed occurrence of target motifs but point towards a higher density of significant matches in the CDSs.
The analysis using a p-value cutoff of 0.01 (Figure 15) showed largely similar results. In RNAcompete and HT-SELEX, both the full transcripts' and the 3'-UTRs' coverages gained in significance.
These results, however, were not reproduced with more stringent cutoffs, as can be seen in Figure 14. This is understandable, as short motifs may not have 1e3 or 1e4 different possible scores and, therefore, cannot produce matches. This fact is illustrated by a reduction in the number of proteins exhibiting matches as the cutoff increases.

What could, however, be observed was a stronger tendency towards outliers as stringency was increased. Since individual proteins are represented by colored dots spreading along the vertical axes of each experiment, this effect is visualized clearly.

3.1.1. Data restriction to single matrix per protein

As a second approach, only one matrix per protein in the datasets was used. This means that, even though some proteins are able to bind more than one type of motif, thus resulting in multiple probability matrices per protein, only one matrix per protein was used in scanning the transcriptome using FIMO. This restriction in data was of interest as it leads to results that remain unbiased in terms of coverage values per protein. As differences in quality scores among the set of matrices available for a single RBP vary only slightly, the one appearing in the last position of each set was chosen for this analysis.

For p-value cutoffs 0.05 and 0.01 (Figure 16), which only include motif-sequence match-scores within the top 5 % or 1 %, respectively, a significant effect could indeed be observed. The coverage values of RNAcompete and HT-SELEX motifs in 3'-UTRs as well as in the full transcripts lie in the 99th percentile of the average coverage that could be observed in the entire transcriptome. In Figure 16, this effect is portrayed by a positive deviation from the horizontal mean-line in any of the subsequence coverages.

3.2. Reproducibility using exact motif-sequence matches

In an exact matching procedure, an alignment of these "best" motifs to every point along a sequence is attempted and exact matches reported. From a probabilistic perspective, motif length plays a major 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. Figure 12 shows a distribution of motif lengths among all experiments.

An analysis of this type was carried out on the same dataset as the previous ones. One difference to remark is in the number of motifs available. While RNAcompete and HT-SELEX offer matrices for each binding motif, this is not the case in SELEX. This dataset includes more motifs than in the previous analysis because a set of similar motifs, for which only one matrix would be available, was listed as multiple individual entries.

4. Discussion

The complementarity hypothesis has led to some insightful findings involving protein-RNA interactions [10] [25] and presents a powerful framework for why these interactions can occur. Using position probability matrices as a mathematical model for the binding affinities of a protein's binding domain, this framework was put to the test in a search for motif occurrences in autogenous transcripts of select RBPs in the human transcriptome using FIMO.

The findings of the analysis carried out here show that RBPs have, on average, higher incidence of binding motifs in their autogenous mRNAs when the cutoff determining significant motif occurrence is not too stringent. This effect is apparent in the variation of a p-value cutoff determining what counts as a significant occurrence: at 0.05 and 0.01, where the top 5 % and top 1 % most specific motif-sequence matches are considered, respectively, enrichment in motif occurrence in autogenous mRNAs was observed. A possible interpretation of this fact may be that energetically weak biases, rather than highly specific forces, are what drive the interactions between proteins and RNAs. These biases include, for example purine density profiles in mRNA and guanine-affinity profiles in proteins [10] and can, on average, cause significant differences in binding behavior.
It is, however, certainly not enough to point to complementarity of physicochemical properties of biological macro-structures for explaining the wide range of protein-RNA interactions out there. The presence of identifiable RNA-binding domains within proteins, some of which bind highly specifically, indicates that it is also a point-wise congruence that drives an RBP to bind a given mRNA. As certain RNA-binding domains are more specific in their choice of partner than others, it is possible that the relatively weak biases throughout the sequences help proteins pre-select their binding partners in the densely populated molecular landscape of the cell? Disentangling these two effects would require an in-depth analysis of complementary binding behavior of multiple RNA-binding domains in different cellular contexts.

As a particularly strong enrichment could be observed in the 3'-UTR and the full transcript, a certain correlation with the average length of the scanned sequences may be suspected. However, without further investigation of this correlation in a statistically sound manner, no

conclusion can be drawn. A comparison between the coverage value of a transcript and its length might be sufficient to uncover this effect. The way PSSMs are created offers an alternative perspective on the matter. A background frequency distribution of the entire transcript (Figure 11) is used in constructing the PSSMs. This, however, contains no information about where in the transcript certain nucleotide combinations are more likely to occur. A higher motif density in this context together with the high average length of the 3'-UTR would directly influence the enrichment of binding in the full transcript. Since the 3'-UTR is known to be involved in post-transcriptional regulation of mRNA [26], a real biological effect may even await discovery at the bottom of this question.

Figure 11 also shows that the 3'-UTR is higher in G and C than the full transcript. In [10], Žagrović and Polyansky mention that guanine and cytosine play a special role in complementarity interactions, thus raising the question whether enrichment in this untranslated region may point to an evolutionarily conserved density of G and C in these contexts.

In conclusion, the complementarity hypothesis offers an exciting new avenue for describing and predicting protein-RNA interactions. Due to the ubiquity of protein-RNA interactions in the cell, their exploration in the context of complementarity may lead to novel insight and may open the door to many new areas of research in biophysics and molecular biology.

5. Figures

5.1. Dataset properties

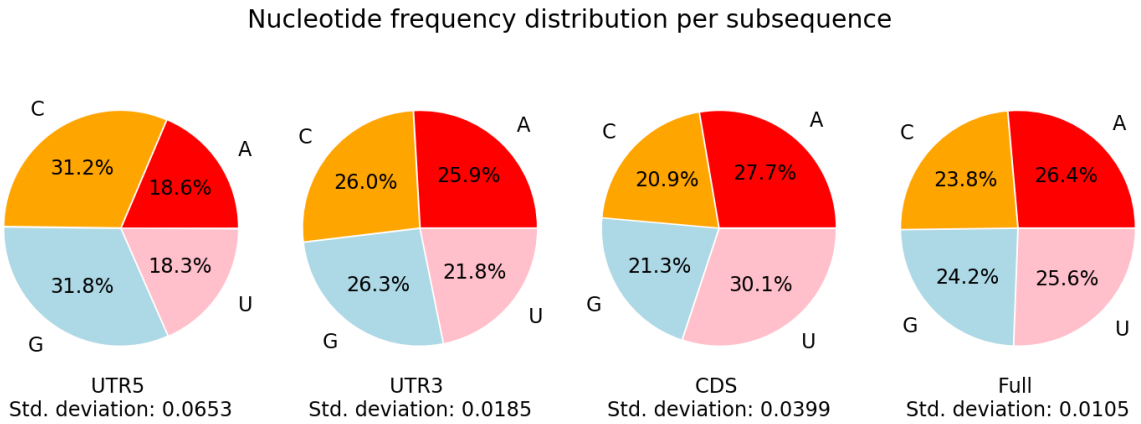

Figure 11: *Nucleotide frequency distribution in the sequences scanned by FIMO.*

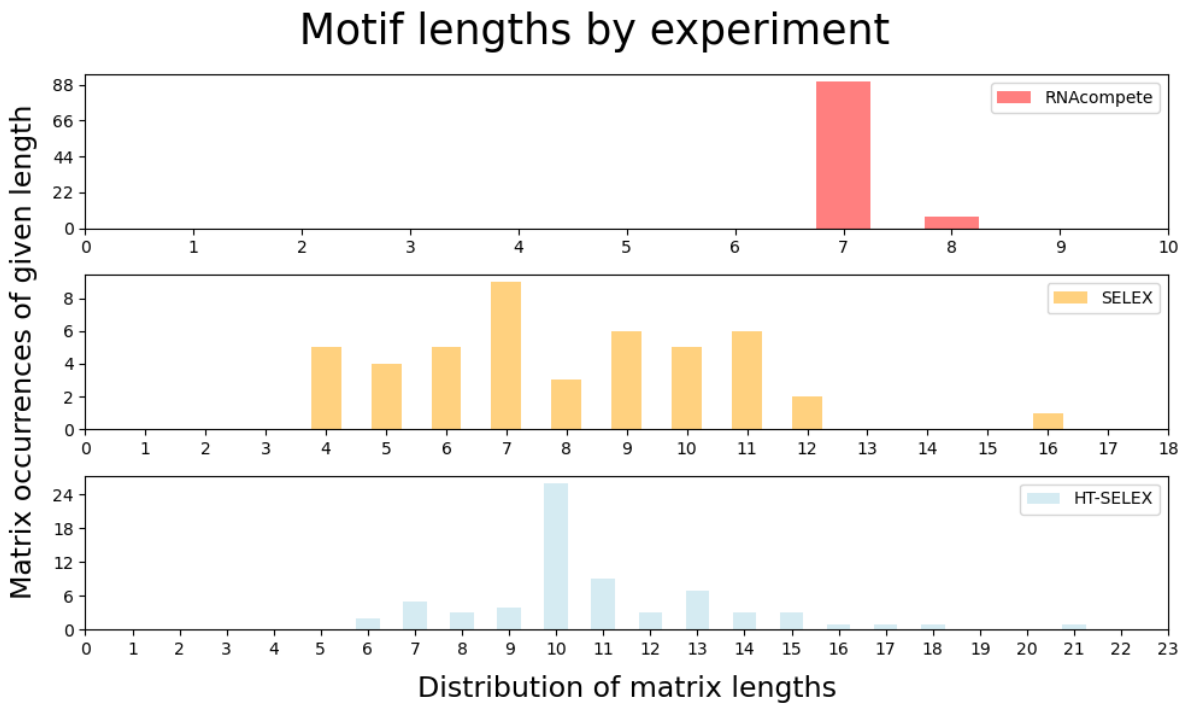

Figure 12: *The y-axes of each subplot are normalized to the total number of matrices of a given length; the x-axes are normalized to the maximum motif length available in the dataset.*

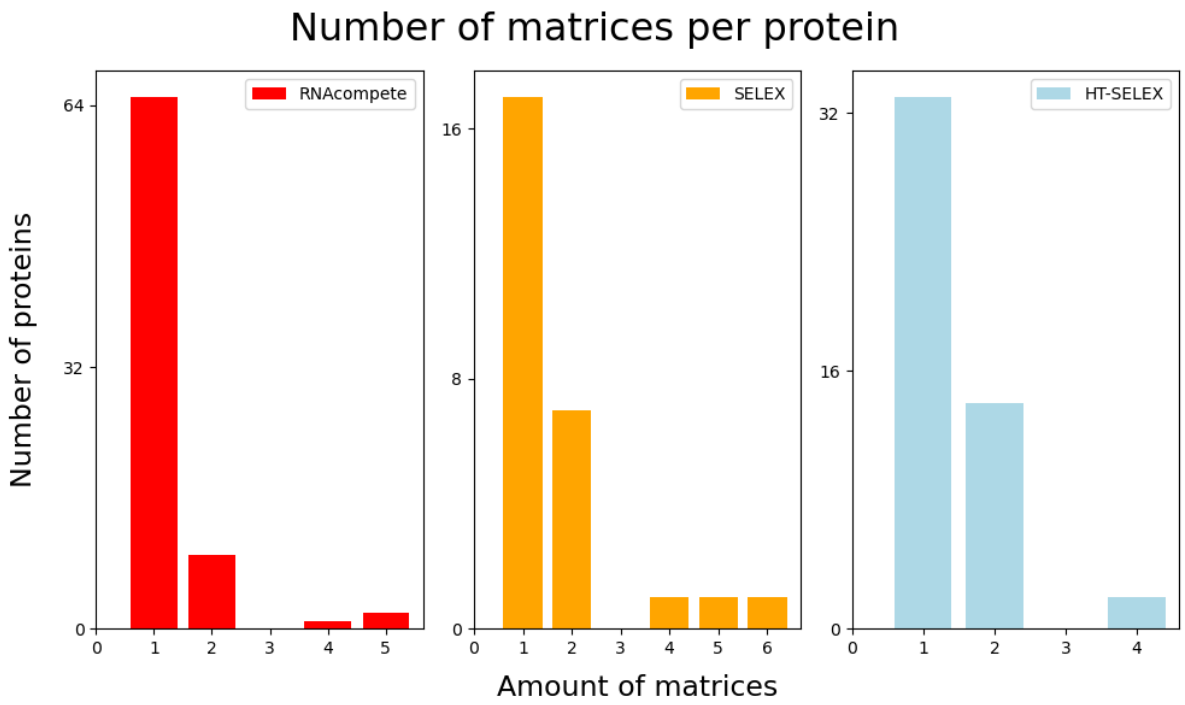

Figure 13: *The y-axis shows the number of proteins in the dataset for which a given number of matrices are available.*

5.2. Normalized analysis using multiple matrices per protein

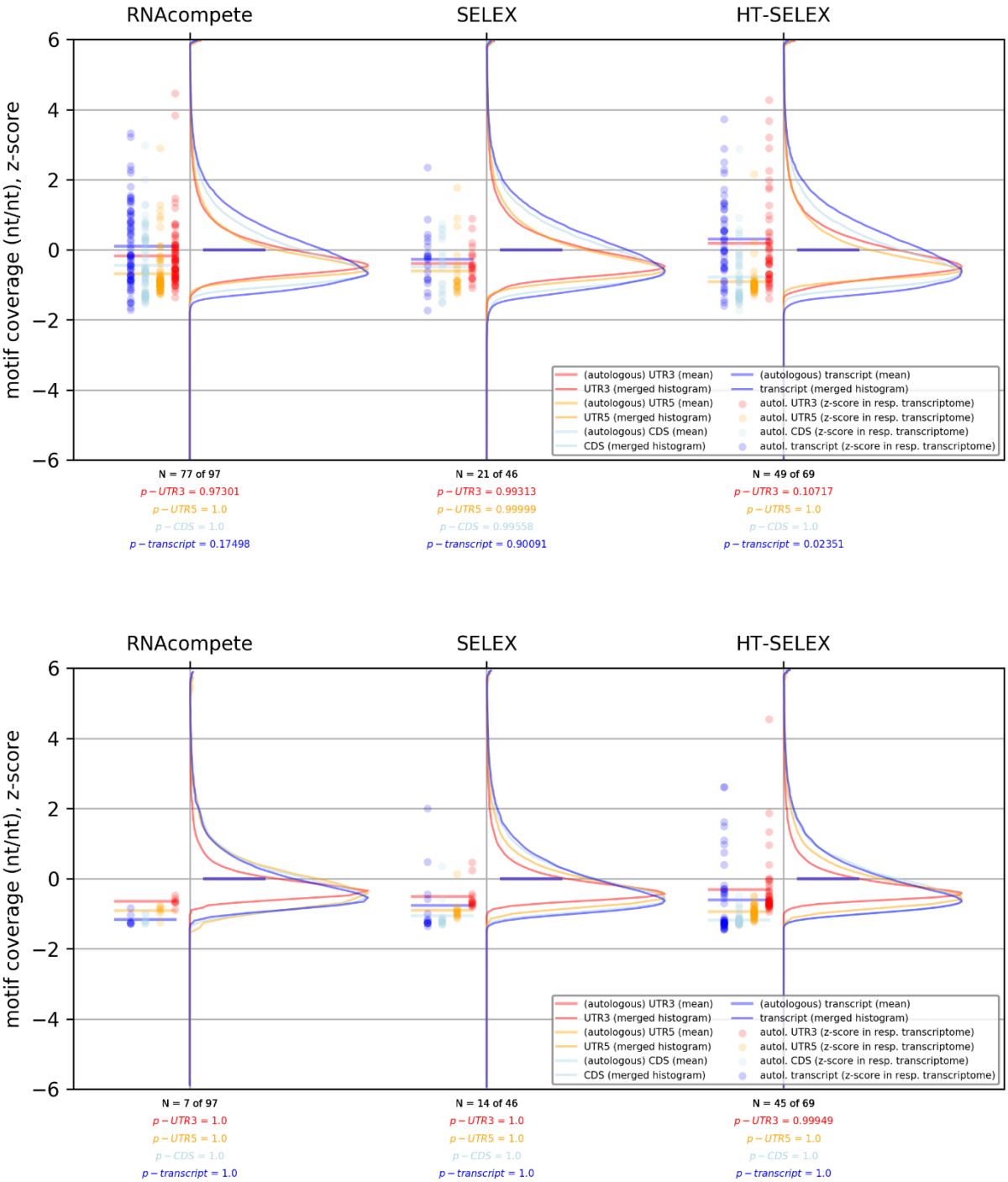

Figure 14: *At a p-value cutoff of 0.001 (top) and 0.0001 (bottom), significant enrichment in autogenous binding could no longer be observed. For each protein, at least one matrix was used to scan for matches.*

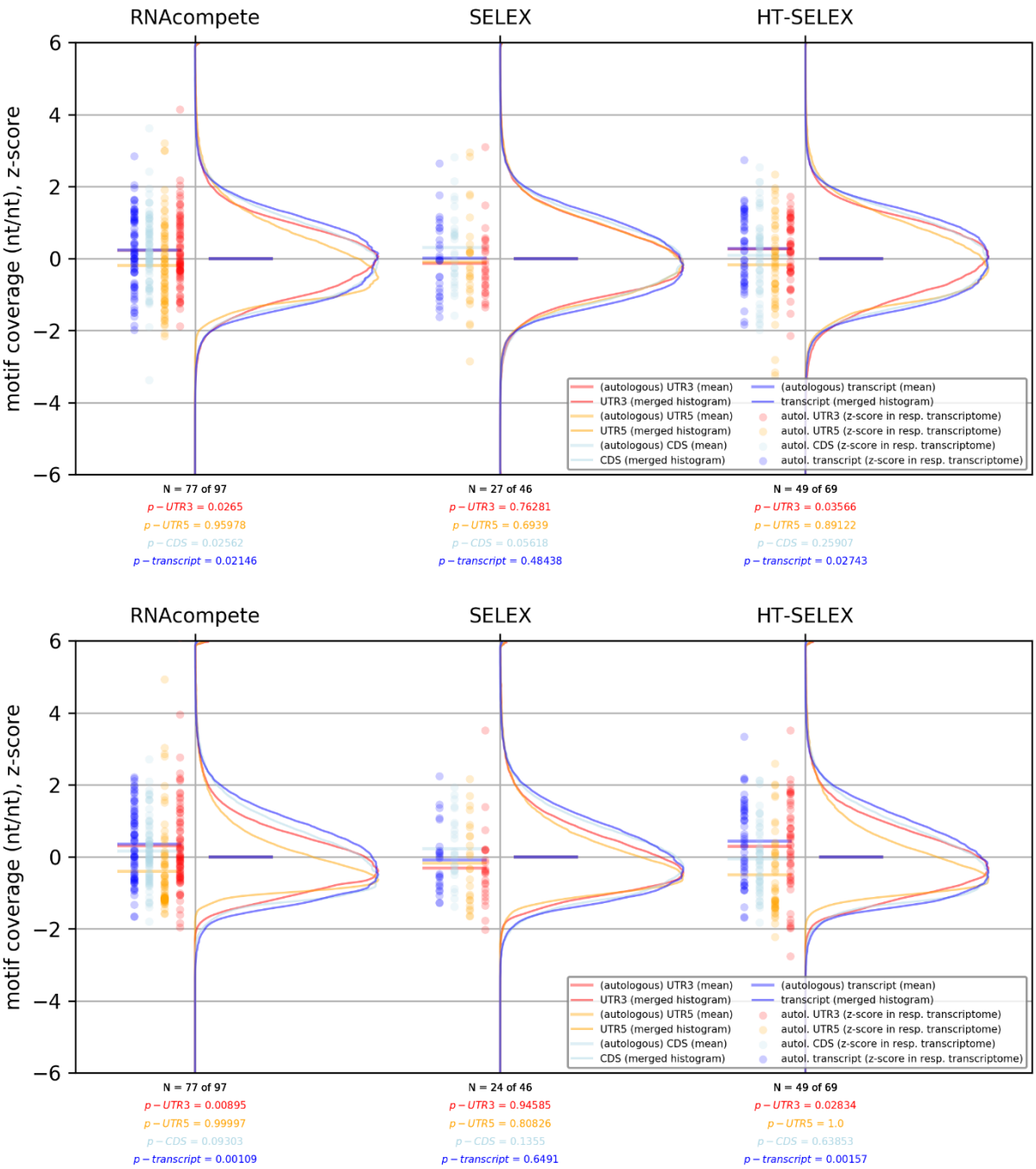

Figure 15: *At a p-value cutoff of 0.05 (top) and 0.01 (bottom), significant enrichment in autogenous binding could be observed. For each protein, at least one matrix was used to scan for matches.*
The vertical axis of this plot shows the distribution of z-scores, while the horizontal axis is separated into the three datasets that were used in the analysis (RNAcompete, SELEX, HT-SELEX). The left side of the graph in each method captures the z-scores of autogenous 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 that 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 autogenous matches' z-scores came from the same distribution as the null-model z-scores.

5.3. Analysis using single matrix per protein

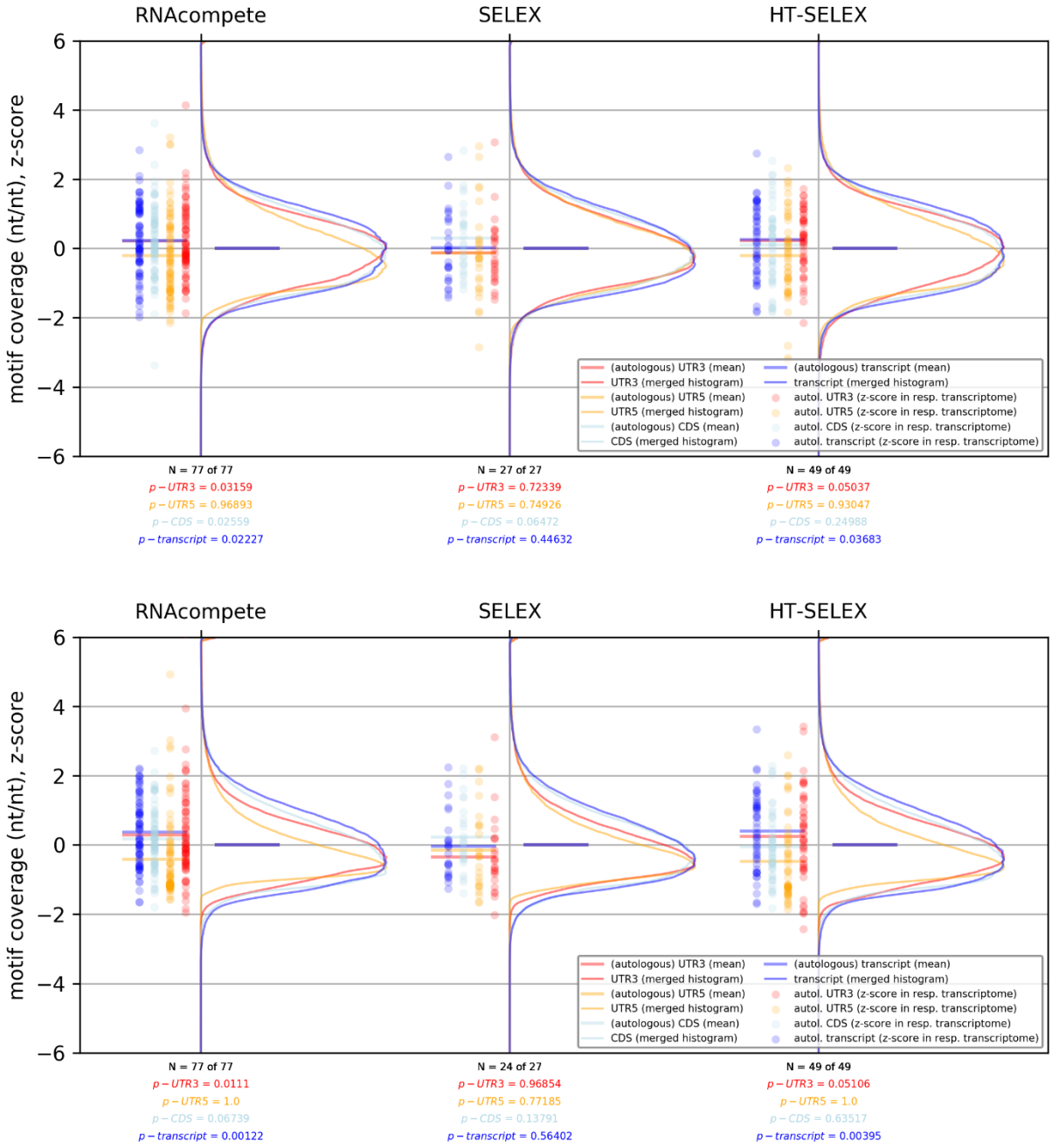

Figure 16: *Analysis of enrichment in autogenous binding using FIMO with a p-value cutoff of 0.05 (top) and 0.01 (bottom). This analysis includes only one matrix per protein.*

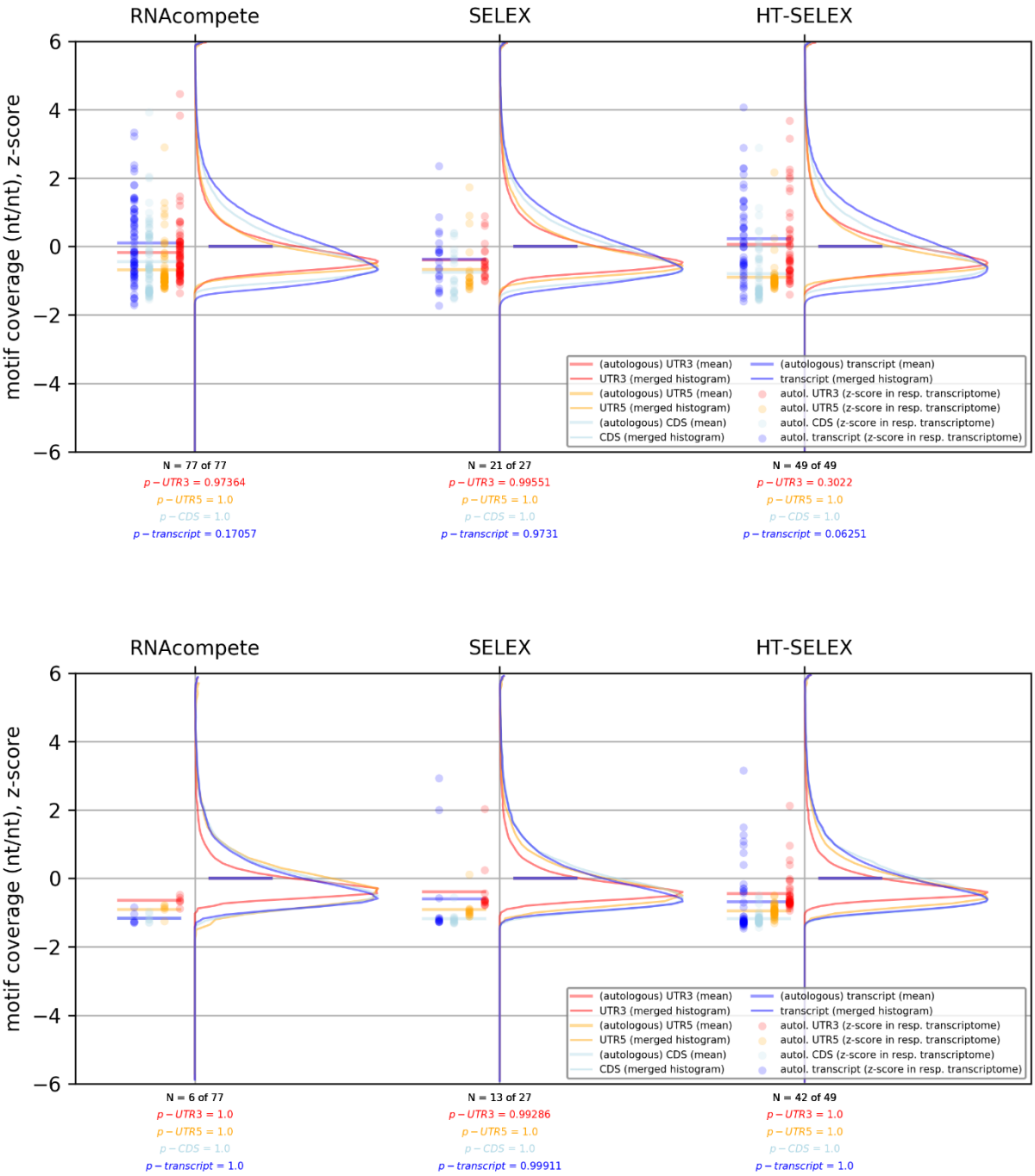

Figure 17: *At a p-value cutoff of 0.001 (top) and 0.0001 (bottom), no significant effect could be observed for an analysis of autogenous transcript motif coverage using a single matrix per protein*

5.1. Reproducibility using exact motif-sequence matches

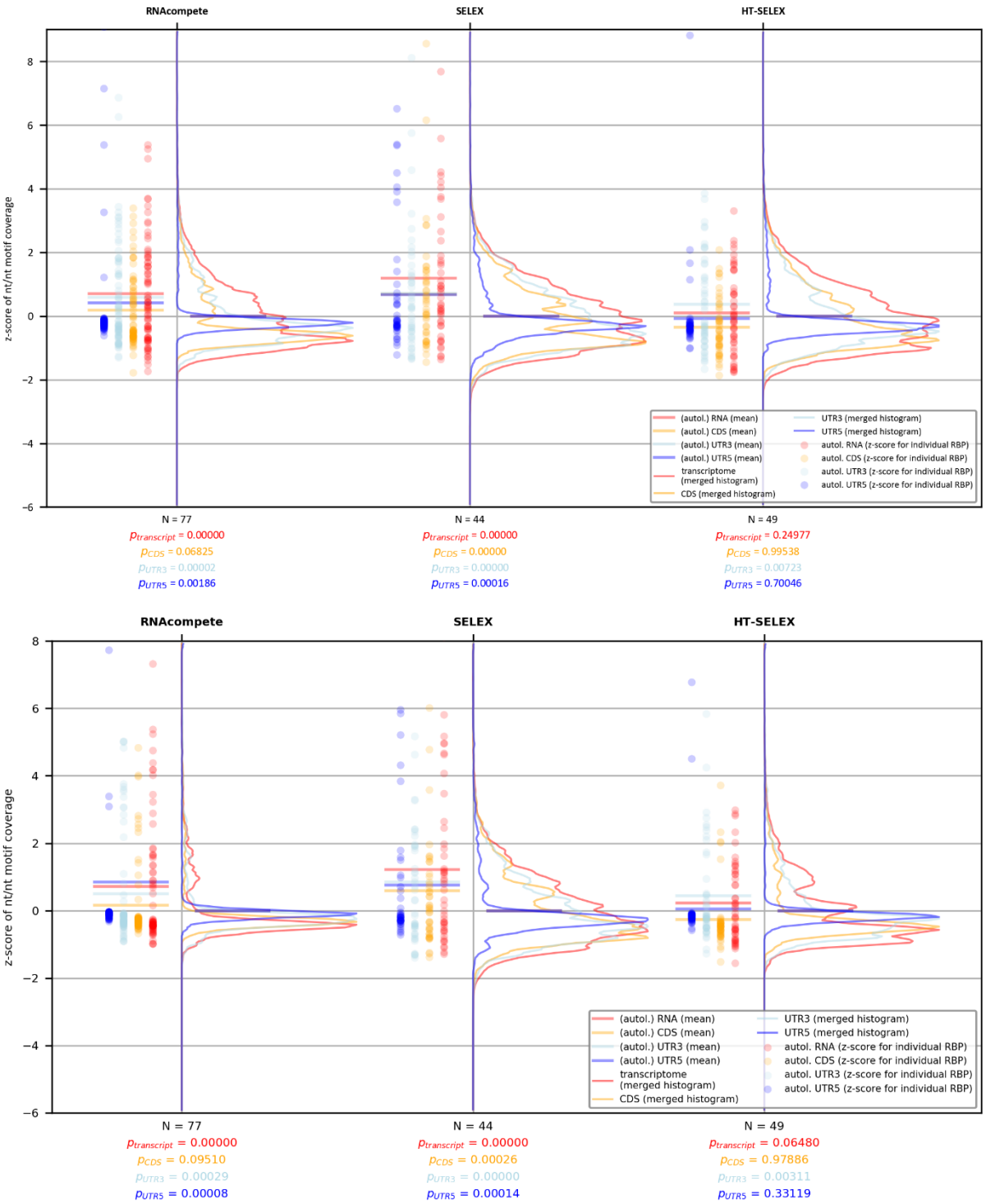

Figure 18: *Exact matching procedure using motif length cutoff of 6nt (top) and 7nt (bottom) with shuffled transcriptome sequences as a null-model. In both analyses, enrichment of autogenous binding could be observed. The CDS shows the highest average density in exact motif-sequence matches.*

6. References

[1] F. Crick, "Central Dogma of Molecular Biology," *Nature,* vol. 227, pp. 561-563, 1970.

[2] M. Müller-McNicoll, O. Rossbach, J. Hui and J. Medenbach, "Auto-regulatory feedback by RNA-binding proteins," *Journal of Molecular Cell Biology,* vol. 11, no. 10, p. 930–939, 2019.

[3] M. Corley, M. C. Burns and G. W. Yeo, "How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms," *Molecular Cell,* vol. 78, no. 1, pp. 9-29, 2020.

[4] R. E. Halbeisen, A. Galgano, T. Scherrer and A. P. Gerber, "Post-transcriptional gene regulation: From genome-wide studies to principles," *Cellular and Molecular Life Sciences,* vol. 65, no. 5, pp. 798-813, 2008.

[5] G. T, B. JL, Y. J and D. G, "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] H. M, P. AA and Z. B, "Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels," *Nucleic Acids Research,* vol. 40, no. 18, pp. 8874-8882, 2012.

[7] Z. B, B. L and P. AA, "RNA-protein interactions in an unstructured context," *FEBS Letters,* vol. 592, no. 17, pp. 2901-2916, 2018.

[8] C. R. Woese, "On the evolution of the genetic code," *PNAS,* vol. 54, no. 6, pp. 1546-1552, 1965.

[9] E. V. Koonin and A. S. Novozhilov, "Origin and Evolution of the Universal Genetic Code," *Annual Reviews of Genetics,* vol. 51, pp. 45 - 62, 2017.

[10] 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.

[11] A. C. Allain and F. H.-T., "From Structure to Function of RNA Binding Domains," in *Madame Curie Bioscience Database [Internet]*, Austin (TX), Landes Bioscience, 2000-2013.

[12] P. A. Chong, R. M. Vernon and J. D. Forman-Kay, "RGG/RG Motif Regions in RNA Binding and Phase Separation," *Journal of Molecular Biology,* vol. 430, no. 23, pp. 4650-5685, 2018.

[13] 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.

[14] R. D, H. KCH, N. K, Z. H, H. TR and M. QD., "RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins," *Methods,* pp. 118-119, 2017.

[15] F. Yasmeen, H. Seo, N. Javaid, M. S. Kim and S. Choi, "Therapeutic Interventions into Innate Immune Diseases by Means of Aptamers," *Pharmaceutics,* vol. 12, no. 10, p. 955, 2020.

[16] C. E. Grant, T. L. Bailey and W. S. Noble, "FIMO: scanning for occurrences of a given motif," *Bioinformatics,* vol. 27, no. 7, pp. 1017-1018, 2011.

[17] G. G, S.-C. F, T. C and L.-P. E, "ATtRACT-a database of RNA-binding proteins and associated motifs," *Database (Oxford),* 2016.

[18] P. Flicek, M. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. Kähäri, S. Keenan, E. Kulesha, F. Martin, T. Maurel, W. McLaren, D. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. Trevanion, A. Vullo, S. Wilder, M. Wilson, A. Zadissa, B. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. Zerbino and S. Searle, "Ensembl 2014," *Nucleic Acids Research,* no. 42, pp. 749-755, 2014.

[19] K. JB, F. JD, B. KA, J.-Z. C, P. VG, L. J, K. K, Z. AM and V. PD, "Xenbase, the Xenopus model organism database; new virtualized system, data types and genomes," *Nucleic Acids Research,* no. 43, pp. 756-763, 2015.

[20] S. N, A. B, A. C, C.-T. A, C. I, G. R, G. N, T. H. P, K. S, L. R, L. W, L. X, L. R, P. N, P. S, P. S, R. R, R. M, S. A, S. D, T. AL, V. D, V. Zalunin and G. Cochrane, "Content discovery and retrieval services at the European Nucleotide Archive," *Nucleic Acids Research,* no. 43, pp. 23-29, 2015.

[21] U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research,* no. 43, pp. 204-212, 2015.

[22] A. Jolma, J. Zhang, E. Mondragón, E. Morgunova, T. Kivioja, K. U. Laverty, Y. Yin, F. Zhu, G. Bourenkov, Q. Morris, T. R. Hughes, L. J. 3. Maher and J. Taipale, "Binding specificities of human RNA-binding proteins toward structures and linear RNA sequences," *Genome Research,* vol. 30, no. 7, pp. 962-973, 2020.

[23] 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.

[24] T. L. Bailey, J. Johnson, C. E. Grant and W. S. Noble, "The MEME Suite," *Nucleic Acids Research,* vol. 43, no. 1, pp. 39-49, 2015.

[25] 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.

[26] E. Matoulkova, B. Vojtesek, E. Michalova and R. Hrstka, "The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells.," *RNA Biology,* vol. 9, no. 5, pp. 563-576, 2012.

[27] A. Theuer, "Analysis of autologous binding preferences of RNA-binding proteins," *Bachelor Thesis,* 2022.

[28] B. Zagrovic, L. Bartonek and A. A. Polyansky, "RNA-protein interactions in an unstructured context," *FEBS letters,* vol. 592, pp. 2901-2916, 2018.

[29] 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.

[30] 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.

[31] 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.

7. Acknowledgements

Many thanks to Prof. Dr. Bojan Žagrović for the amazing opportunity of joining the Žagrović research group for the short duration of this project and for lending invaluable support whenever necessary, while also giving me the space to explore this engaging research topic on my own. Being welcomed into the research group enabled me to work in a stimulating environment full of challenges to overcome and insights to gain.

I want to thank my direct supervisor, Thomas Kapral's, who's help was essential during the time I spent on this project. Not only did he lay the groundwork for this investigation, his technical know-how and scientific rigor played a vital role in progressing my work at a constant rate, and his openness for questions and willingness to help gave me the support needed to solve even the most difficult problems.

A big thank you also to Arthur Theuer for laying a solid foundation that I could build upon and for providing me with the code I could use to reproduce my result using the exact-matching procedure.

8. 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