Identification and Analysis of Anchor Residues within Transcription Factor Binding Sites (TFBS)

Riddhi Goswami* J. Angel Arul Jothi* Debostuti Ghoshdastidar†

Abstract

This research project is dedicated to advancing the field of drug design by identifying and analysing anchor residues within transcription factor binding sites (TFBS). We achieve this through a comprehensive examination of UniProbe data and the application of MEME Suite tools. Our ultimate goal is to enable the development of novel therapeutic strategies that inhibit gene expression in diseased states by targeting these critical residues. Given data availability quality and computational resources constraints, we have adopted a systematic approach to data procurement preparation and analysis. The initial steps involve collecting and preprocessing relevant data from UniProbe, primarily focusing on the human transcription factors, GATA4 and ETV5. Subsequently, the project outlines the process of motif identification using MEME Suite's command-line interface, interpreting generated Probability Weight Matrices (PWM) and applying a consensus approach. We introduce a novel binning strategy to discern anchor residues by comparing motif prevalence across various data segments, thereby identifying pivotal anchor residues. Our meticulous approach instils confidence in the reliability of our research.

Furthermore, we propose an algorithm to align datasets, which is a step that could potentially transform the field of drug design. This methodology provides a deeper understanding of the complex mechanisms of TFBS, laying the groundwork for developing drugs that can modulate gene expression by precisely targeting anchor residues. The outcomes of this project will be a comprehensive report on the methodology findings and implications supported by visual representations of the data and adaptable code for broader dataset application. Through this initiative, we aim to significantly contribute to the rapidly evolving field of targeted drug design, which offers a promising path for treating various genetic disorders. We believe that our project has the potential to inspire and motivate further research in this area.

Keywords: Transcription Factor Binding Sites (TFBS); Anchor Residues; Probability Weight Matrix (PWM); Sequence Analysis; Computational Biology.

1 Introduction

In the realm of targeted drug design, a select few drugs have been developed with the precision to inhibit Transcription Factor (TF) binding sites, thereby preventing undesired gene expression. This approach represents a frontier in therapeutic development, offering the potential to modulate disease-related genes directly. Our research, by introducing a novel methodology to predict anchor residues within TF binding sites, aims to enhance the precision of targeted drug design significantly. This could have a profound impact on modulating gene expression in disease states. However, challenges persist in accurately predicting the binding sites due to the complex dynamics of protein-DNA interactions.

Techniques such as targeted protein degradation, as presented by Dai et al. [1], leverage the proteasome's ability to degrade specific tumour transcription factors, thus preventing them from binding to DNA. Inamoto and Shin [2] explore peptide therapeutics that mimic or block TF binding sites, providing a direct mechanism for inhibiting transcription factor activity. Bushweller's [3] work on transcription factors highlights the strategic targeting of specific domains within these proteins, effectively turning 'undruggable' targets into viable ones. Lambert et al. [4] discuss small molecule inhibitors that disrupt the interaction between transcription factors and DNA by identifying critical residues essential for binding. Gayvert and Elemento [5] employ computational methods to repurpose existing drugs by predicting their impact on transcription factor activity. They highlight the potential of in silico approaches in identifying

*Department of Computer Science, BITS Pilani Dubai
†Department of Biotechnology, BITS Pilani Dubai

drug candidates with the desired effect on TFs. However, accurately predicting the binding sites of transcription factors remains a significant challenge. Advanced computational models like RosettaFold [6] and AlphaFold [7] have marked substantial progress in predicting protein structures, yet translating these structures into predictable TF binding dynamics encompasses a myriad of complexities, underscoring the intricate nature of the problem.

Against this backdrop, our research proposes three promising methods to predict anchor residues within TF binding sites leveraging identified motifs from the Universal PBM Resource for Oligonucleotide Binding Evaluation (UniPROBE). The research is currently carried out on the transcription factors, GATA4 and ETV5. The first method utilises MEME Suite's [13] tools for motif identification, on a binning strategy to identify critical anchor residues. The second method builds upon this and applies calculated weights on the generated Position Weight Matrix (PWM) for each bin. The final approach employs aligning the entire dataset acquired from UniPROBE alongside the top binding sequence.

This paper is divided into eight sections. Section 1 introduces the study's objectives and significance in targeted drug design. Section 2 reviews existing literature on computational biology advancements related to TFBS prediction. Section 3 describes the dataset used, focusing on the UniProbe database. Section 4 outlines the tools employed, particularly the MEME Suite for motif analysis. Section 5 details the methodology behind the three approaches to identify anchor residues. Section 6 discusses implementation details, specifying the technical setup and software used. Section 7 presents the results and discussion, analysing the findings and their implications. Finally, Section 8 concludes the study, summarising the outcomes and proposing directions for future research.

2 Literature Review

The field of computational biology has undergone a profound transformation, particularly in predicting and understanding complex biological structures. These revolutionary advances are pivotal for decoding the myriad interactions that underpin biological functions and therapeutic interventions and inspire optimism about the field's future. This literature review delves into recent seminal works that epitomise the progress in this arena, highlighting methodologies, achievements, and their broader implications.

2.1 Accurate Prediction of Protein-Nucleic Acid Complexes Using RoseTTAFoldNA

One of the landmark studies in computational biology is the introduction of RoseTTAFoldNA, a novel machine-learning approach designed to enhance the prediction accuracy of protein-nucleic acid complexes significantly. Building upon the foundational RoseTTAFold method, this technique represents a significant stride towards understanding complex biological structures. Its improved accuracy over prior models, including AlphaFold, guarantees the model's reliability. RoseTTAFoldNA employs a sophisticated deep-learning architecture.

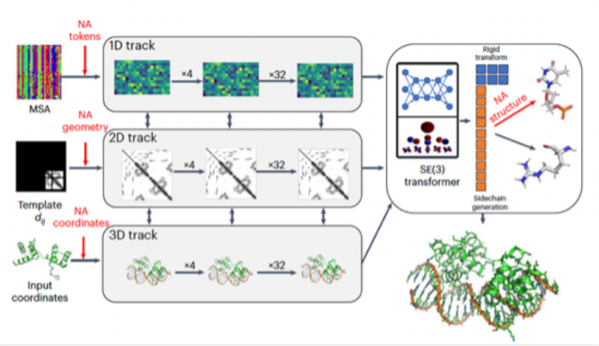


Figure 1: Overview of the architecture of RoseTTAFoldNA

RoseTTAFoldNA is trained on a diverse dataset from the Protein Data Bank (PDB) that includes protein monomers, complexes, RNA monomers, dimers, and protein-RNA/DNA complexes, utilising a sophisticated architecture to achieve its predictive accuracy. This model is notably enhanced by integrating ten extra tokens for the four DNA and RNA nucleotides, unknown DNA and RNA, and the 22 tokens for amino acids in the 1D track. This enhancement significantly improves the model's capability to process and recognise nucleic acids. Furthermore, the generalisation of the 2D track to include interactions between nucleic acid bases and between bases and amino acids enables the model to capture the complex interplay within protein-nucleic acid complexes. In the 3D track, the detailed representation of nucleotides, which extends to accurately constructing all atoms in the nucleotide, mirrors the comprehensive approach used for amino acids. The entire architecture is supported by a network of 36 three-track layers, further augmented by four structure refinement layers. The model's complexity is evident in its 67 million parameters, which are crucial for its high predictive accuracy and performance.

This comprehensive approach, enabled by the potential of deep learning, allows for the accurate modelling of intricate protein-nucleic acid interactions. The model's superior performance in predicting protein-nucleic acid complex structures, surpassing other state-of-the-art methods, sets a new benchmark in computational biology. It underscores the exciting potential of deep learning in unravelling and understanding the complexities of biology.

2.2 Highly Accurate Protein Structure Prediction with AlphaFold

The advent of AlphaFold by DeepMind marks a paradigm shift in protein structure prediction. AlphaFold achieves unprecedented accuracy in predicting protein structures through a sophisticated deep-learning model, solving a decades-old challenge in the field. This development significantly enhances our understanding of protein folding mechanisms, a fundamental aspect of biological research. It opens new avenues for further exploration and discovery in this field and pharmaceutical development.

The AlphaFold model, heralding a monumental leap in protein structure prediction, was meticulously trained on a comprehensive dataset of multiple sequence alignments (MSAs) and structural data derived from various protein databases. This rigorous training process, initially harnessing the Uniclust dataset, which contains 355,993 sequences with full MSAs, followed by a final training phase that applied identical hyperparameters predominantly using data from the Uniclust prediction set and, to a lesser extent, a clustered set from the Protein Data Bank (PDB), instils confidence in the reliability of AlphaFold's predictions.

The Big Fantastic Database (BFD)—a vast, publicly accessible compilation of protein families used by numerous CASP teams—played a critical role in this process. It offers an extensive resource of 65,983,866 families represented through MSAs and hidden Markov models covering over 2.2 billion protein sequences from various sources, including reference databases, metagenomes, and metatranscriptomes. This comprehensive data source ensures AlphaFold is well-informed and equipped for accurate protein structure prediction.

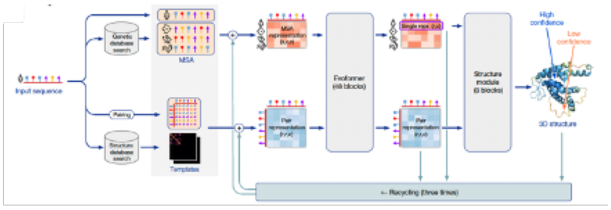

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Figure 2: Model architecture of AlphaFold

AlphaFold's innovative architecture and methodologies mark significant departures from traditional models, impressively employing a deep learning neural network designed for high precision without the need for ensembling, thus improving prediction speeds significantly. An attention-based neural network is central to its design, allowing for the discernment of long-range interactions between amino acids. This is complemented by the integration of evolutionary information through MSAs, aiding in the prediction of accurate residue contacts and structure determinations. Spatial graph convolutional networks are employed to model the geometric relationships between amino acids accurately, and an end-to-end prediction with a subsequent refinement process ensures the model's predictions closely align with actual structures. Using frame-aligned point error (FAPE) further underscores the model's capability to accurately align predicted points with their proper positions, showcasing AlphaFold's unparalleled precision and profound impact on computational biology.

AlphaFold's accuracy and efficiency in protein structure prediction testify to AI's potential in revolutionising computational biology. Its ability to predict near-experimental accuracy structures not only paves the way for groundbreaking research in understanding biological processes and drug discovery but also instils hope for future applications in other areas of computational biology.

2.3 Identification of Human Lineage-Specific Transcriptional Coregulators

This section reviews the paper on identifying human lineage-specific transcriptional coregulators, highlighting the development of GENRE and a glossary of TF-8mer modules. The study introduced a glossary of 108 non-redundant TF-8mer modules derived from protein-binding microarray data encompassing 671 metazoan TFs.

These modules are instrumental in deciphering the specificity shared by transcription factors, allowing researchers to pinpoint both direct and indirect TF binding sites. The precision of these 8mer modules was rigorously compared against traditional position weight matrices. Utilising 239 ENCODE TFhostilein immunoprecipitation sequencing datasets and associated RNA sequencing profiles, the 8mer modules demonstrated a superior capacity in identifying indirect binding motifs and their tethering transcription factors. A key innovation presented in the paper is GENRE (Genomically Equivalent Negative Regions), a tool designed to construct matched genomic background sequences for regulatory region analysis. GENRE addresses a critical challenge in regulatory genomic studies by providing a tunable approach to create negative control regions that closely mirror the genomic areas under study. The tool's effectiveness surpassed four other state-of-the-art methods previously used for background sequence construction. This capability enhances the reliability of genomic studies by ensuring that the background sequences used in comparative analyses are truly representative.

These studies underscore the rapid advancements in computational biology, each contributing unique insights into biological systems' structural and regulatory complexities. These works pave the way for future research through innovative methodologies and deep learning approaches, promising significant impacts on biological discovery and therapeutic development.

3 Dataset Description

UniPROBE is a comprehensive platform that meticulously collates and shares data on protein-DNA interactions. At its core are insights from universal Protein-Binding Microarray (PBM) experiments, which reveal the binding affinities of proteins—primarily TFs—across all conceivable DNA sequences of a given length, known as k-mers.

Acquired datasets include files containing the transcription factor's preferred binding site sequences, presented in either consensus format or as PWMs. Additionally, there is microarray data, encompassing raw and processed results from PBM experiments that reveal fluorescence intensities, indicative of the strength of interaction between the transcription factor and various DNA sequences. Enrichment scores (E-Score) within the dataset provide a comparative analysis of a sequence's binding affinity to the transcription factor.

Among the many files available for TFs, `..._8mers_top_enrichment.txt` was used in our evaluation. This file involves analysis of 8-mer sequences (sequences consisting of 8 nucleotides) and their binding affinities to the TF, as measured in PBM experiment. Here is what each column represents:

- **8-mer**: This column lists the 8-nucleotide sequences that were tested for binding affinity with the protein.
- **8-mer (Complement)**: Reverse complement, with A pairing with T and C pairing with G, read in the opposite direction.
- **E-score**: The enrichment score quantifies the binding affinity to the specific 8-mer sequence relative to a background or control. A positive E-score indicates a higher affinity, whereas a negative E-score suggests a lower affinity than the background. This score helps identify preferred binding sequences.
- **Median**: This column represents the median fluorescence intensity from the PBM experiments for each 8-mer sequence. It provides a central tendency measure of the binding signal, which helps assess the strength of interaction between TF and the DNA sequence.
- **Z-score**: The Z-score standardises the binding affinity measurement, indicating how many standard deviations an observation is from the mean. A positive Z-score means the observation is above the mean, while a negative Z-score indicates it is below the mean. This helps identify sequences with significantly higher or lower binding affinities.

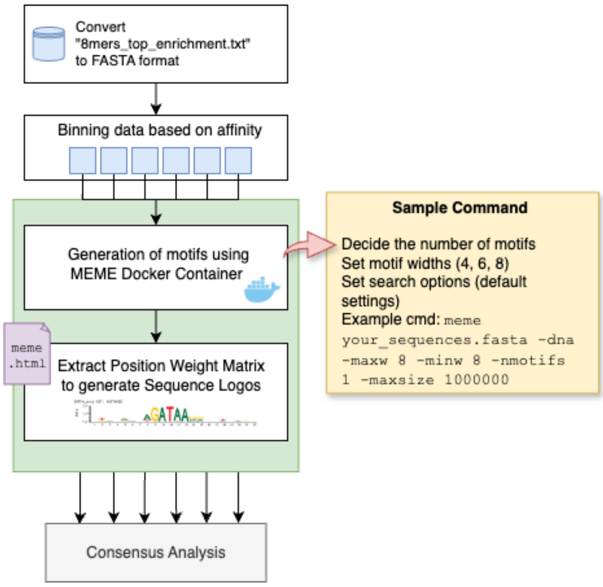
Notably, this file focuses on sequences with the highest enrichment scores, indicating they are among the most preferred binding sites for TF. Including sequences with dots (e.g., A..AGATAAG) suggests a tolerance for mismatches or gaps in specific positions within the 8-mer, potentially indicating flexibility in the binding site preference of TF. This information is crucial for understanding the specific DNA sequences TF targets, which can inform studies on gene regulation, transcription factor networks, and the molecular mechanisms underlying TF's roles in development and disease.

4 Methodology

Three methods were devised, described in detail in this section.

4.1 Method 1

Before detailing the methodology, the architecture diagram (Figure 3) describes the process in brief.


Convert
"8mers_top_enrichment.txt"
to FASTA format
Binning data based on affinity
Generation of motifs using
MEME Docker Container
meme
.html
Extract Position Weight Matrix
to generate Sequence Logos
GATAA
Consensus Analysis
Sample Command
Decide the number of motifs
Set motif widths (4, 6, 8)
Set search options (default
settings)
Example cmd:

```
meme
your_sequences.fasta -dna
-maxw 8 -minw 8 -nmotifs
1 -maxsize 1000000
```

Figure 3: Architecture Diagram of Method 1

The chosen `...._8mers_top_enrichment.txt` file from UniPROBE was first converted into FASTA format using a Python script. After which, the sequences were categorised into discrete bins representing ten per cent intervals with no overlap.

This method uses MEME Suite, an integral suite of bioinformatics tools dedicated to discovering and analysing motifs within biological sequences, catering to protein and nucleic acid (DNA/RNA) data. Among the many tools it offers, Multiple EM for Motif Elicitation (MEME) was used as it identifies statistically significant, repeated patterns (motifs) within sequences and helps discover novel motifs in related sequences.

4.1.1 MEME Configuration

A Python script was developed to use the official MEME Docker container to automate the feeding of bin-specific FASTA data into the MEME command-line interface. The generated output files for each bin were placed into corresponding bin directories to prevent overwriting.

Some input parameters are required to configure MEME. These parameters include deciding on the motif width, where the minimum and maximum widths of the motifs of interest must be determined; default settings can serve as a guideline for those uncertain about this parameter. A range of 4, 6, and 8 were considered the possible motif widths of interest. The number of motifs to be identified by MEME is another crucial input. Starting with a smaller number is recommended for initial explorations to focus on the most significant motifs (a value of 1 was chosen to get the most considerable motif). Additionally, search options must be selected to dictate the frequency with which MEME should identify motifs within the sequences—ranging from motifs that occur just once per sequence, zero or one time, to motifs that may appear multiple times. Typically, the default search option suffices for most analyses, providing a balanced approach to motif discovery. To implement in the command line, a sample command is:

```
meme your_sequences.fasta -dna -maxw 8 -minw 8 -nmotifs 1 -maxsize 1000000
```

This instruction tells MEME to seek only one motif within the DNA sequences, with motif widths spanning eight nucleotides, within a dataset optimised for analysis efficiency.

4.1.2 Consensus Analysis

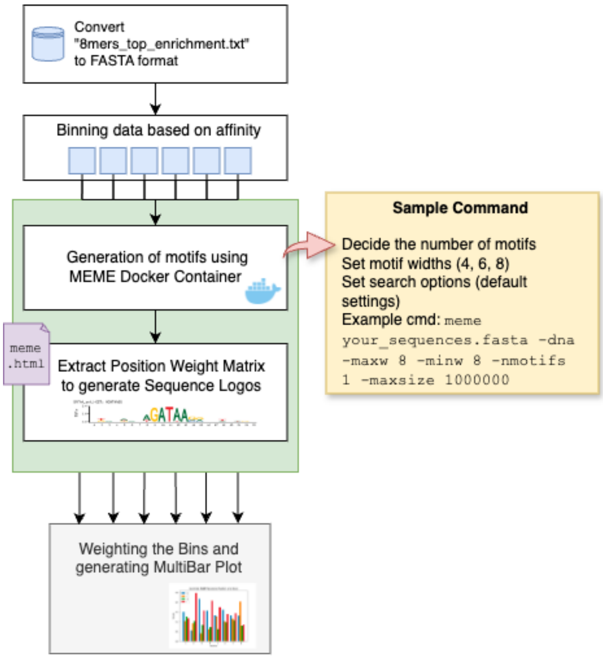
Among the genertaed output files, the file meme.html is the most significant as it encapsulates the motif discovery results. A script was developed to extract the PWM from a div tag within this file. The PWM is essentially a matrix where the columns represent the four bases (A, C, G, T) and the rows correspond to the length of the extracted motif, as specified by the user. The extracted PWMs are then analysed using a consensus approach [14]. In this analysis, the frequency of each base at each position within the motif for each bin is recorded, and the base with the highest overall frequency at each position is noted. It is used to determine the most prevalent residue at each position within a PWM and a consensus sequence is generated to highlight critical binding affinities.

4.1.3 Problems Associated

The primary issue with this implementation course is that the position of the anchor residue has yet to be determined. This issue arises because the generated motifs are of different lengths, and where they occur in the 8mer sequence is unknown.

4.2 Method 2

This approach builds upon the previous method but in addition, it assigns weight to the PWMs of each bin to get an understanding of the probability of each base in each position. The disadvantage of the previous method is that the PWMs generated for different lengths are not super-imposable that is the motifs acquired are different from each other so the position of the motif concerning the entire sequence will be difficult to derive. This method looks promising in getting an understanding of which bases are present throughout the dataset and at what position will be achieved. The below architecture diagram (Figure 4) describes the procedure in brief.


Convert "8mers_top_enrichment.txt" to FASTA format
Binning data based on affinity
Generation of motifs using MEME Docker Container
meme .html
Extract Position Weight Matrix to generate Sequence Logos
GATAA
Weighting the Bins and generating MultiBar Plot

Sample Command

Decide the number of motifs
Set motif widths (4, 6, 8)
Set search options (default settings)
Example cmd: `meme your_sequences.fasta -dna -maxw 8 -minw 8 -nmotifs 1 -maxsize 1000000`

Figure 4: Architecture Diagram of Method 2

4.2.1 Assigning Weights

Here, the issue arises of how to choose the weights. Each bin contains 10 per cent of the dataset without overlap, and the E-score varies unevenly in each bin. As a result, a metric is chosen to calculate based on the E-Score values of the bins. As a result, the weights assigned were based on the average and standard deviation of the E-Score values of the sequences present.

The **metric** for each bin i is calculated as:

$$\text{metric}_i = \frac{\text{avg}_i}{\text{std}_i}$$

The **weight** for each bin i is calculated as:

$$\text{weight}_i = \frac{\text{metric}_i}{\sum_j \text{metric}_j}$$

where,

$$\begin{aligned} \text{metric}_i &: \text{Metric for bin } i \\ \text{avg}_i &: \text{Average E-score value for bin } i \\ \text{std}_i &: \text{Standard deviation E-score value for bin } i \\ \sum_j \text{metric}_j &: \text{Sum of all metrics} \end{aligned}$$

These weights were multiplied with the 8mer PWMs generated, and a bar chart was plotted. The rows denoted the eight positions with 4 bars denoting the four bases (A, C, G, T) for each position. The columns denote the probability of each base's binding affinity (the higher the probability - the more likely the base is significant for binding). In this manner, the bases with the highest probability along with their position are acquired.

4.3 Method 3: BINGO Algorithm

The final method is the most promising approach yet. This method is different from the previous methods explained so far. This method doesn't use binning and instead aligns alongside the top motif, the entire dataset `..._8mers_top_enrichment.txt` in FASTA format. The below architecture diagram (Figure 5) describes the process in brief:

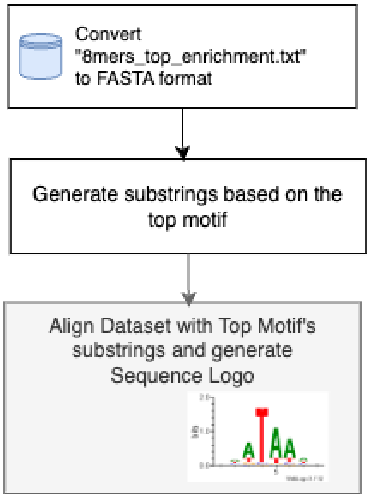

Convert "8mers_top_enrichment.txt" to FASTA format
Generate substrings based on the top motif
Align Dataset with Top Motif's substrings and generate Sequence Logo

Figure 5: Architecture Diagram of Method 3

4.3.1 Generating Substrings

The top sequence is taken as this has the highest binding affinity. For this 8mer, the script generates all possible substrings. These substrings range from a length of 1 base to 8 bases, and are derived from the original 8mer sequence. The generated substrings are ordered in a list lexicographically and numerically, meaning that longer substrings come first in the list, and substrings of the same length are ordered alphabetically. Below is an example of how to generate and order such substrings:

1. Given an 8mer sequence, such as `ATCGTAGC`.
2. The script generates all possible substrings:
   - 1-base substrings: `A`, `T`, `C`, `G`, `T`, `A`, `G`, `C`
   - 2-base substrings: `AT`, `TC`, `CG`, `GT`, `TA`, `AG`, `GC`
   - 3-base substrings: `ATC`, `TCG`, `CGT`, `GTA`, `TAG`, `AGC`
   - 4-base substrings: `ATCG`, `TCGT`, `CGTA`, `GTAG`, `TAGC`
   - 5-base substrings: `ATCGT`, `TCGTA`, `CGTAG`, `GTAGC`
   - 6-base substrings: `ATCGTA`, `TCGTAG`, `CGTAGC`
   - 7-base substrings: `ATCGTAG`, `TCGTAGC`
   - 8-base substring: `ATCGTAGC`
3. The substrings are then ordered lexicographically and numerically:
   - 8-base substring: `ATCGTAGC`
   - 7-base substrings: `ATCGTAG`, `TCGTAGC`
   - 6-base substrings: `ATCGTA`, `TCGTAG`, `CGTAGC`
   - 5-base substrings: `ATCGT`, `TCGTA`, `CGTAG`, `GTAGC`
   - 4-base substrings: `ATCG`, `TCGT`, `CGTA`, `GTAG`, `TAGC`
   - 3-base substrings: `ATC`, `TCG`, `CGT`, `GTA`, `TAG`, `AGC`
   - 2-base substrings: `AT`, `TC`, `CG`, `GT`, `TA`, `AG`, `GC`
   - 1-base substrings: `A`, `T`, `C`, `G`, `T`, `A`, `G`, `C`

This ordered list of substrings can be used for further analysis, ensuring that longer substrings and those appearing earlier in the alphabet are prioritized.

4.3.2 Aligning the Substrings

The remaining sequences in the dataset will be aligned based on the longest substring present in each sequence, and these will be aligned against the top motif. Subsequent sequences are then aligned based on the longest possible substring they share with the top 8mer. This method identifies the sequence segment that is most frequently occurring throughout the dataset. Consequently, the anchor residue and its position relative to the top 8mer are determined. This below example illustrates how the sequences are aligned based on the longest common substrings with the top 8mer:

1. Consider the following sequences, where `AGATAAGG` is the top sequence:

   ```
   AGATAAGG
   CTTATCGC
   AGATAACG
   AAGATAAG
   ```

2. The remaining sequences are aligned alongside the top motif based on the longest common substrings:

   ```
   AGATAAGG
   -CTTATCG
   AGATAACG
   AGATAAG-
   ```

4.3.3 Algorithm

The following is the algorithm for the entire approach discussed above:

Algorithm 1 Align Sequences to Motif

```
function ReadSequencesFromFile($filePath$)
    $sequences \leftarrow []$
    for $line$ in open($filePath$) do
        if $line[0] \neq '>'$ then
            $sequences$.append($line$.strip())
        end if
    end for
    return $sequences$
end function

function RefSeq($sequences$)
    for $sequence$ in $sequences$ do
        return $sequence$
    end for
    return $None$
end function

function GenerateSubstrings($baseStr$)
    $substrings \leftarrow []$
    $length \leftarrow len(baseStr)$
    for $i \leftarrow 0$ to $length - 1$ do
        for $j \leftarrow i + 1$ to $length$ do
            $substrings$.append($baseStr[i : j]$)
        end for
    end for
    return $sorted(set(substrings), key = len, reverse = True)$
end function

function AlignSequences($sequences$, $referenceSequence$, $motif$)
    $substrings \leftarrow$ GenerateSubstrings($motif$)
    $df \leftarrow$ DataFrame(columns=[$-7$ to 14])
    for $sequence$ in $sequences$ do
        $sequence \leftarrow sequence$.replace('.', ”)
        $aligned \leftarrow False$
        $entire \leftarrow False$
        $idx \leftarrow sequence$.find($motif$)
        if $idx \neq -1$ then
            $aligned \leftarrow True$
            $entire \leftarrow True$
        else
            for $substring$ in $substrings$ do
                if $substring$ in $sequence$ then
                    $idx \leftarrow sequence$.find($substring$)
                    $refIndex \leftarrow referenceSequence$.find($substring$)
                    $aligned \leftarrow True$
                    $entire \leftarrow False$
                    break
                end if
            end for
        end if
        if $aligned$ then
            $offset \leftarrow refIndex - idx$
            $indices \leftarrow range(offset, offset + len(sequence))$
            $df$.append(DataSeries($sequence$, index=$indices$))
        end if
    end for
    return $df$
end function
```

5 Implementation Details

All the scripts were developed and run on MacBook Pro (14-in, 2021), Chip: Apple M1 Pro (8-core CPU, 14-core GPU), Memory: 16 GB, macOS: Sonoma 14.5. Software requirements: Python 3.9.13, Docker version 25.0.3, build 4debf41, memesuite/memesuite: Docker image of MEME Suite

6 Results and Discussion

The discussed methods were applied to two well-known Transcription Factors, GATA4 and ETV5. The results derived were cross-checked and verified experimentally by performing protein crystallisation.

6.1 PWM from UniProbe

PWMs are a crucial tool for understanding transcription factor binding sites. It represents the frequency of each nucleotide at every position in a binding site. This subsection discusses the PWMs derived from PBM experiments for both GATA4 and ETV5, sourced from the UniProbe database. PBM is an in vitro high-throughput technique that allows for the determination of binding specificities of transcription factors by measuring their binding to a large array of synthetic DNA sequences. The Seed-And-Wobble method is often employed in these experiments to fine-tune the binding specificity measurements.

Below are the PWM representations for GATA4 and ETV5 transcription factors derived from PBM experiments:

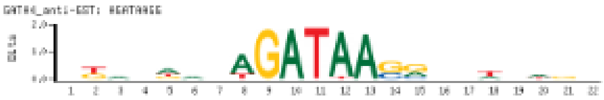


Figure 6: Binding Site Motif for GATA4

Figure 6 illustrates the PWM for GATA4. This motif is derived from a Seed-And-Wobble PBM experiment, showcasing the nucleotide preferences at each position within the binding site.

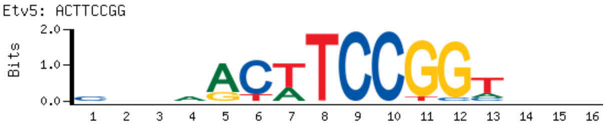


Figure 7: Binding Site Motif for ETV5

Similarly, Figure 7 shows the PWM for ETV5, another transcription factor analyzed using the Seed-And-Wobble method in PBM experiments. The PWM reflects the sequence specificity of ETV5, highlighting the most preferred nucleotides at each position.

The graphical representations of these PWMs offer a visual summary of binding preferences, making it easier to interpret and compare transcription factor binding specificities.

6.2 Results from Method 1

In method 1, the discovered motifs were analysed using consensus sequence to characterise DNA-binding motifs across discrete 10 per cent bins of the entire dataset. The below PWMs represent the results procured through MEME for discrete 10 per cent bins. The change in motif length has caused the differences in PWMs.

6.2.1 GATA4 Result discussion

For the 4mers (Figure 8), the consensus sequences ranged from GATA in bins 1-5 to variations like ATC, TTGC, and CGCG in the subsequent bins, with GATA being the most recurrent. The frequency analysis of individual positions within these motifs revealed A and T as the predominant nucleotides in the second and third positions, respectively, indicating their potential significance in the binding specificity of these sequences.

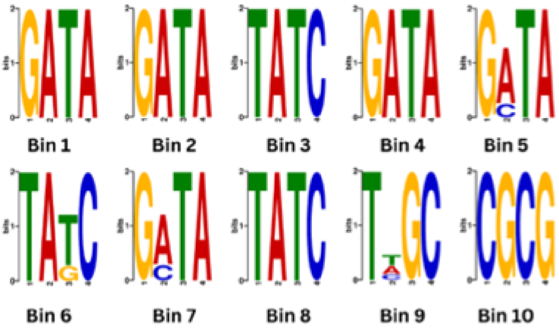


Figure 8: GATA4 - Motifs of Length 4 Identified in Each Discrete 10% Bin

Expanding the analysis to 6mer motifs (Figure 9), the diversity among consensus sequences became more pronounced, ranging from CTTATC to ATCTAA. This analysis underscored A and T as the most common nucleotides at the sixth and second positions, respectively, suggesting a preference for these bases in the binding sites for transcription factors.

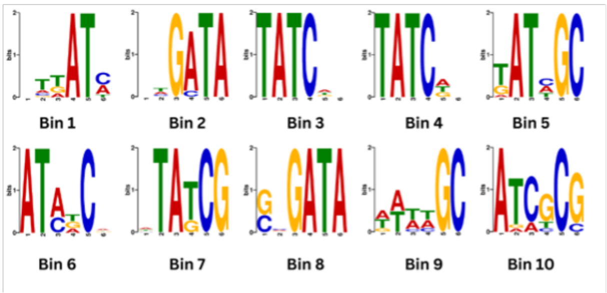


Figure 9: GATA4 - Motifs of Length 6 Identified in Each Discrete 10% Bins

Similarly, for 8mer motifs (Figure 10), consensus sequences exhibited even more significant variability, with TTATAACA and GATCAGAT bookending the range of sequences found. Again, A and T dominated, underscoring their prevalence in these longer motifs. As per the binning methodology, the possible anchor residue is A or T due to its widespread occurrence across various motif lengths and bins.

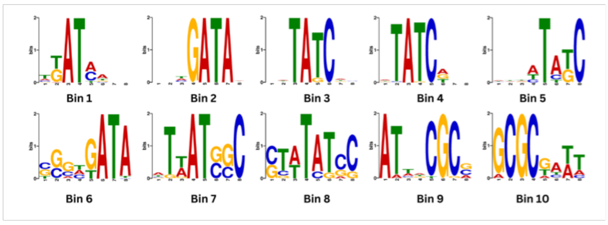


Figure 10: GATA4 - Motifs of Length 8 Identified in Each Discrete 10% Bin

So, it can be derived that the motif GATA or its reverse complement TATC is consistently identified across most bins, indicating accurate motif detection.

6.2.2 ETV5 Result Discussion

For the 4mer motifs (Figure 11), the motifs ranged from TCCG to GGAA with the binning, with GGAA being the most recurrent. The frequency analysis of individual positions within these motifs revealed T in the first position, G in the second position, C in the third position, and A in the fourth position as

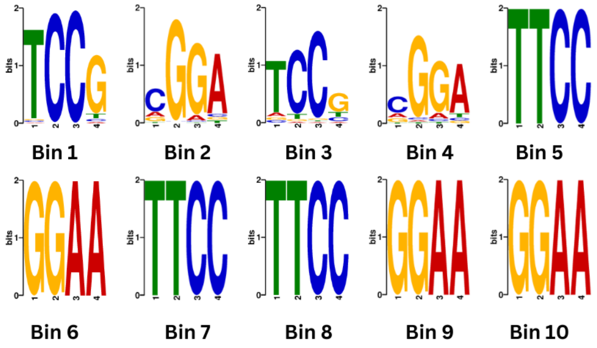


Figure 11: ETV5 - Motifs of Length 4 Identified in Each Discrete 10% Bin

the predominant nucleotides. This suggests that these nucleotides play significant roles in the binding specificity of these sequences.

Expanding the analysis to 6mer motifs (Figure 12), the diversity among motifs became more pronounced, ranging from CCGGAA to GGAAAA. This analysis underscored G at the third and fourth positions, and A at the fifth and sixth positions as the most recurrent. These findings highlight the preference for these bases in the binding sites for transcription factors.

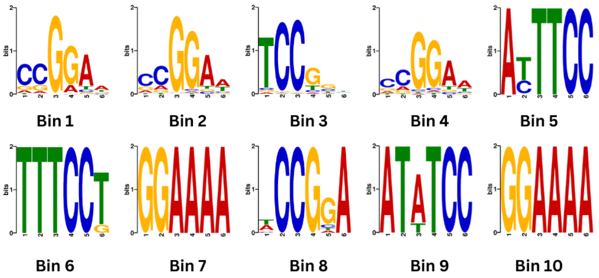


Figure 12: ETV5 - Motifs of Length 6 Identified in Each Discrete 10% Bin

Similarly, for 8mer motifs (Figure 13), motifs exhibited significant variability, starting with CCGGAAGT and TATTACCG bookending the range of sequences found. A and G dominated many positions, particularly A, which appeared most frequently at positions 1, 2, 4, 5, 7, and 8. However, the most frequent was A at the eighth position with eight out of ten occurrences. This prevalence underscores their significance in these longer motifs. As per the binning methodology, the possible anchor residue is likely A in the eighth due to its widespread occurrence across various motif lengths and bins.

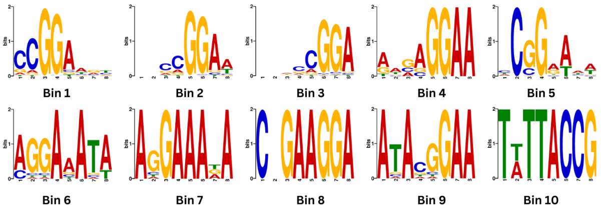


Figure 13: ETV5 - Motifs of Length 8 Identified in Each Discrete 10% Bin

Finally, it can be derived that the motif GGAA or its reverse complement TTCC is consistently identified across most bins, indicating accurate motif detection.

However, the problem with this method is that it is difficult to confirm the possible position of the anchor residues. Each motif varies based on length, and bin distribution and it's difficult to unanimously get the position of the anchor residues.

6.3 Results from Method 2

In method 2, the results were diagrammatically plotted on a bar chart. The rows denoted the eight positions with 4 bars denoting the four bases (A, C, G, T) for each position. The columns denote the probability of each base's binding affinity (the higher the probability - the more likely the base is significant for binding). In this manner, the bases with the highest probability along with their position are acquired.

6.3.1 GATA4 Result Discussion

Based on Figure 14, it can be derived that T in the second position is a possible anchor residue. However, upon closer inspection, it can be seen that the binding affinity is just less than 0.5, which is not a clear majority. Here the concept of threshold needs to be addressed. It is not right to simply take the highest anchor residue but instead, a threshold must be considered. Without further data, no such calculation could be made. However, compared to method 1, at least we get a clear picture of the binding affinity of each base in each position.

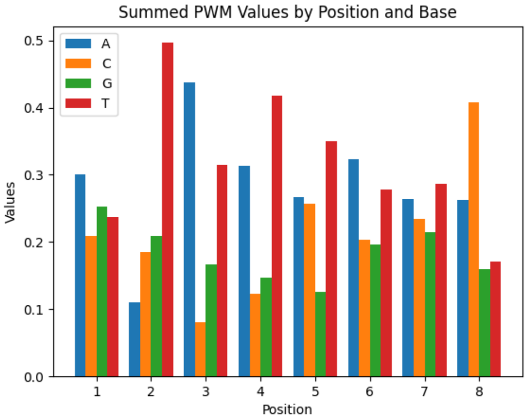


Figure 14: GATA4 - Summed PWM Values as per Weighted Bins

6.3.2 ETV5 Result Discussion

Similarly, based on Figure 15, it can be derived that A in the eighth position is a possible anchor residue. However, unlike in GATA4, here the the binding affinity is about 0.7, which is a much higher majority. G in the third position is the second highest with a binding affinity close to 0.6, and may also be considered. Because of this ambiguity in where to draw the line on possible anchor residue, this method was not pursued further. However, there is a future possibility of coming up with a threshold value.

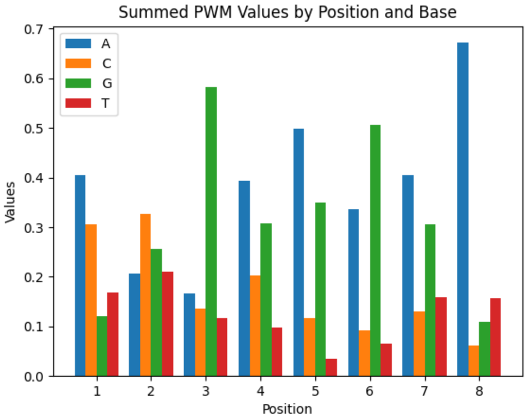


Figure 15: ETV5 - Summed PWM Values as per Weighted Bins

6.4 Results from Method 3

The final method is the most promising approach with a clear anchor residue being found as a result of producing a sequence logo. After aligning the entire dataset with respect to the highest binding affinity motif, the derived result will have no ambiguity concerning the position of the anchor residue.

6.4.1 GATA4 Result Discussion

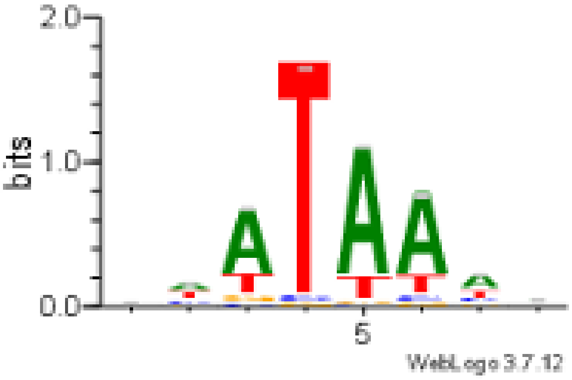


Figure 16: GATA4 - Sequence Logo after Aligning Entire Dataset

Based on the sequence logo generated (Figure 16), it can concluded that T at position 4 is the most prevalent across the entire dataset of GATA.

6.4.2 ETV5 Result Discussion

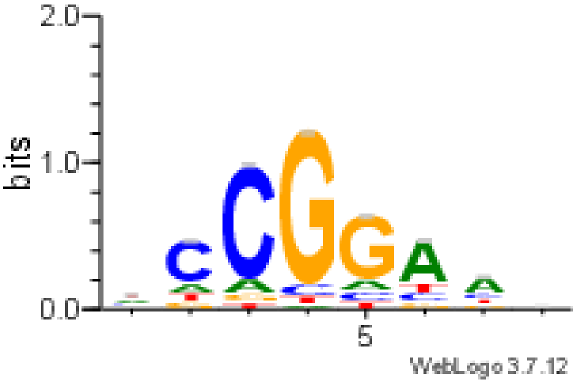


Figure 17: Sequence Logo after Aligning Entire Dataset

Similarly, for ETV5's sequence logo (Figure 17), we can conclude that C at position 3, and G at position 4 are the most prevalent across the entire dataset of ETV5.

7 Conclusion

Identifying the anchor residue likely to be A or T due to its widespread occurrence across various motif lengths and bins presents a challenge in pinpointing its exact position, necessitating further investigation. Understanding the placement of these anchors is crucial. One approach to tackle this issue is aligning the entire dataset per the top motif presented in the project. This methodology was applied across the ETS as a whole domain of proteins, and the results are corroborated with crystal research-derived anchor residues. After analysing the results, the robustness of the algorithm can be determined. Furthermore, anchor residues may only sometimes be sequential and could be spaced intermittently. Employing machine learning [17] or other predictive methods could offer a sophisticated avenue for determining the locations of these anchor residues within the motifs.

References

[1] Dai, M.-Y., Radhakrishnan, S., Li, R., Tan, R., Yan, K., Fan, G., & Liu, M. (2022). Targeted Protein Degradation: An Important Tool for Drug Discovery for "Undruggable" Tumor Transcription Factors. *Technology in Cancer Research & Treatment.* https://doi.org/10.1177/15330338221095950

[2] Inamoto, I., & Shin, J. A. (2019). Peptide therapeutics that directly target transcription factors. *Biopolymers.* https://doi.org/10.1002/PEP2.24048

[3] Bushweller, J. H. (2019). Targeting transcription factors in cancer - from undruggable to reality. *Nature Reviews Cancer.* https://doi.org/10.1038/S41568-019-0196-7

[4] Lambert, M., Jambon, S., Depauw, S., & David-Cordonnier, M.-H. (2018). Targeting Transcription Factors for Cancer Treatment. *Molecules*, 23(6), 1479. https://doi.org/10.3390/MOLECULES23061479

[5] Gayvert, K., & Elemento, O. (2019). Drug-Induced Expression-Based Computational Repurposing of Small Molecules Affecting Transcription Factor Activity. In *Methods of Molecular Biology.* https://doi.org/10.1007/978-1-4939-8955-3_10

[6] Baek, M., McHugh, R., Anishchenko, I., Baker, D., & DiMaio, F. (2022). Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *bioRxiv.* https://doi.org/10.1101/2022.09.09.507333

[7] Jumper, J. M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature.* https://doi.org/10.1038/s41586-021-03819-2

[8] Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A., & Bulyk, M. L. (2017). Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Systems.* https://doi.org/10.1016/J.CELS.2017.06.015

[9] Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(Database). https://doi.org/10.1093/nar/gkn660

[10] Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. (2014). UniPROBE update 2015: New tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 43(D1). https://doi.org/10.1093/nar/gku1045

[11] Beer, S., Björk, M., & Beardall, J. (2021). Fluorescence Measurement Techniques. In *Photosynthesis in Algae: Biochemical and Physiological Mechanisms.* https://doi.org/10.1007/978-981-15-5354-7_26

[12] Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.

[13] Nystrom, S. L., & McKay, D. J. (2021). Memes: A motif analysis environment in R using tools from the MEME Suite. *PLOS Computational Biology*, 17(9), e1008991. https://doi.org/10.1371/journal.pcbi.1008991

[14] Rajamani, D., Thiel, S., Vajda, S., & Camacho, C. J. (2004). Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), 11287–11292. https://doi.org/10.1073/pnas.0401942101

[15] Zhang, J. (2002). Analysis of information content for biological sequences. *Journal of Computational Biology*, 9(6), 885–896. https://doi.org/10.1089/106652702760138583

[16] Zhang, L., & Xiao, F. (2022). Belief Kullback-Leibler Divergence-based Dynamical Complexity Analysis for Biological Systems. In *Proceedings of the IEEE International Conference on Systems Man and Cybernetics* (pp. 1-6). IEEE. https://doi.org/10.1109/ISCTech58360.2022.00009

[17] Ding, P., Wang, Y., Zhang, X., Gao, X., Liu, G., & Yu, B. (2023). DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Briefings in Bioinformatics.* https://doi.org/10.1093/bib/bbad231

[18] Aziz, F., & Al-Rashid, S. Z. (2022). The SVM algorithm predicts DNA binding sites bound to specific transcription factors. *Iraqi Journal of Science*, 63(11). https://doi.org/10.24996/ijs.2022.63.11.37

[19] Wang, W., Jiao, X., Sun, B.-H., Liang, S., Wang, X., & Zhou, Y. (2022). DeepGenBind: A novel deep learning model for predicting transcription factor binding sites. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/BIBM55620.2022.9994984