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

Riddhi Goswami* Dr. Angel Arul Jothi* Dr. 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 factor GATA4. 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; UniProbe Database; MEME Suite; Drug Design; Gene Expression Inhibition; Data Procurement and Preparation; Probability Weight Matrix (PWM); Motif Identification; Disease State Gene Regulation; Computational Biology; Bioinformatics; GATA4 Transcription Factor; Sequence Analysis.

1 Introduction

In the realm of targeted drug design, a select few drugs have been developed with the precision to inhibit Transcription Factor binding sites, thereby preventing undesired gene expression. This approach represents a frontier in therapeutic development, offering the potential to modulate disease-related genes directly. By introducing a novel methodology to predict anchor residues within TF binding sites, our research 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

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

impact on transcription factor activity. They highlight the potential of in silico approaches in identifying 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 AlphaFold and RosettaFold have made 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 introduces a novel methodology to predict anchor residues within TF binding sites leveraging identified motifs from the UniProbe database, a comprehensive resource for transcription factor binding site information. By focusing on the transcription factor GATA4, we utilise the MEME Suite's tools for motif identification, employing a unique binning strategy to highlight critical anchor residues. This approach is predicated on the notion that identifying these anchor residues can significantly enhance the precision of targeted drug design aiming to modulate gene expression in disease states effectively. Our study navigates through the intricacies.

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.

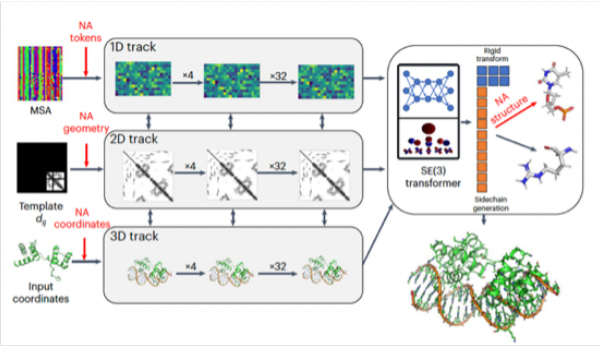

Figure 1: Overview of the architecture of RoseTTAFoldNA

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.

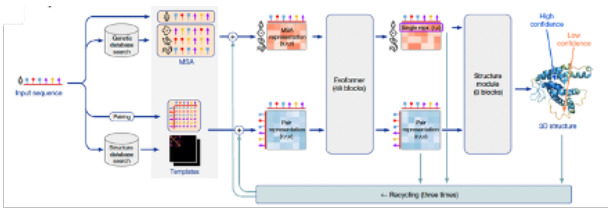

Figure 2: Model architecture of AlphaFold

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

The UniPROBE database [9] for Universal PBM Resource for Oligonucleotide Binding Evaluation 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 transcription factors (TFs)—across all conceivable DNA sequences of a given length known as k-mers. This rich data is instrumental in deciphering gene regulation mechanisms, given the pivotal role of transcription factors in modulating gene expression.

The dataset is a treasure trove of files indispensable for understanding the intricate world of transcription factor-DNA interactions, a key to unravelling the mysteries of gene regulation.

This includes files containing the transcription factor's preferred binding site sequences, presented in either consensus format or as Position Weight Matrices (PWMs), which elucidate the nucleotide preferences at each position within the binding site. Additionally, the dataset contains microarray data encompassing raw and processed results from Protein Binding Microarray (PBM) experiments that reveal fluorescence intensities indicative of the strength of interaction between the transcription factor and various DNA sequences. Enrichment scores within the dataset provide a comparative analysis of a sequence's binding affinity to the transcription factor, facilitating the identification of sequences with high binding preferences. Quality control (QC) files are also included to ensure the integrity of the data, documenting the reproducibility and reliability of the experiments through metrics like correlation coefficients between replicates. Furthermore, analysis reports offer comprehensive summaries of the experimental findings, accentuating high-confidence binding motifs, drawing comparisons with known motifs, and offering insights into the transcription factor's novel binding preferences.

By housing this diverse and detailed dataset, UniPROBE is a crucial resource for the scientific community, fostering a deeper understanding of gene regulation by studying transcription factor-DNA interactions. Here are the descriptions of the files received when choosing the GATA4 transcription factor:

- **GATA4_anti-GST_8mers_11111111.txt**: The sequence of 1s in the file name might indicate a specific condition or filter applied during the analysis, such as a threshold for considering a sequence as a binding site. This file involves analysis of 8-mer sequences (sequences consisting of 8 nucleotides) and their binding affinities to the GATA4 transcription factor as measured in a protein binding microarray (PBM) experiment. The columns represent:
  - *8-mer*: Lists the 8-nucleotide sequences tested for binding affinity with the GATA4 protein.
  - *8-mer (Complement)*: Reverse complement with A pairing with T and C pairing with G, read in the opposite direction.
  - *E-score*: Quantifies the binding affinity of GATA4 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.
  - *Median*: Represents the median fluorescence intensity from the PBM experiments for each 8-mer sequence.

  - *Z-score*: Standardises the binding affinity measurement, indicating how many standard deviations an observation is from the mean.

- **GATA4_anti-GST_8mers_top_enrichment.txt**: This file presents the top enriched 8-mer sequences based on their affinity for binding to the GATA4 transcription factor as determined by protein binding microarray (PBM) experiments. It has the same attributes as the previous file. Notably, this file focuses on sequences with the highest enrichment scores, indicating they are among the most preferred binding sites for GATA4. 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 GATA4.

- **GATA4_anti-GST_alldata.txt**: This file includes detailed experimental data for each microarray spot, including control and experimental probes.

  - *Column & Row*: These columns likely indicate the probe's physical location on the microarray chip.
  - *Name & ID*: Identifiers for each probe or control spot on the array.
  - *Sequence*: The DNA sequence associated with each probe.
  - *Cy3 & Alexa488*: Fluorescence intensity measurements for each probe detected by different fluorophores (Cy3 and Alexa488).
  - *Cy3Flags & Alexa488Flags*: Flags indicating the quality or reliability of the fluorescence measurements.
  - *Cy3Exp*: Expected values or an experimental condition related to Cy3 fluorescence.
  - *Obs/Exp*: Comparison of the fluorescence intensity to an expected value or model prediction.
  - *Alexa488Norm*: Normalised Alexa488 fluorescence intensity.
  - *Alexa488Median*: The median value of Alexa488 fluorescence across multiple measurements or experiments.
  - *Alexa488Adjusted*: Adjusted Alexa488 fluorescence intensity, correcting for background noise or other experimental factors.

- **GATA4_anti-GST_combinatorial.txt**: Represents the results of combinatorial analysis from a PBM experiment involving the GATA4 transcription factor.

- **GATA4_anti-GST_RC.pwm**: Provides the Position Weight Matrix (PWM) for the transcription factor GATA4 based on a specific seed k-mer.

- **GATA4_anti-GST_regression.txt**: The output from a regression analysis modelling the relationship between different 4-mer DNA sequences and their binding affinities to the GATA4 transcription factor.

4 Tools Used

This study utilised MEME Suite [13], an integral suite of bioinformatics tools dedicated to discovering and analysing motifs within biological sequences, catering to protein and nucleic acid (DNA/RNA) data. Its application spans a broad spectrum of genomic and proteomic research areas, including identifying and characterising transcription factor binding sites. Major components of this suite include:

- **Multiple EM for Motif Elicitation (MEME)**: This tool identifies statistically significant, repeated patterns (motifs) within sequences and helps discover novel motifs in related sequences.

- **TOMTOM**: Used to compare discovered motifs against a database of known motifs to find matches and assess their statistical significance, aiding in identifying motifs that resemble known binding preferences.

Among the many tools available, the study has conducted motif discovery and analysis using the MEME tool to investigate and delineate DNA-binding motifs associated with transcription factors, such as GATA4. After identifying the motif, a theoretical framework is used to determine the anchor residue.

5 Methodology

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

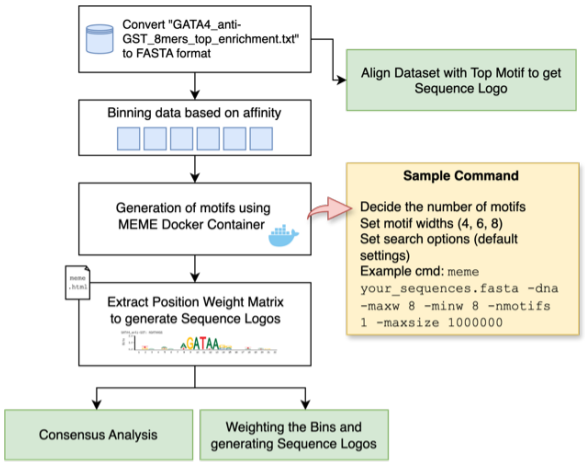

Figure 3: Architecture Diagram

5.1 Method 1

The `GATA4_anti-GST_8mers_top_enrichment.txt` file from the UniProbe dataset was chosen and converted into FASTA format, which MEME requires for processing. In this file, the sequences are ordered based on their binding affinity to GATA4 (the higher the value of the E-score, the better the binding affinity). The sequences were then categorised into discrete bins representing ten percent intervals.

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.

The file meme.html is the most significant among the generated outputs as it encapsulates the motif discovery results. A script was developed to extract the Position Weight Matrix (PWM) from a div tag within this file. The extracted PWMs were then subjected to a consensus approach analysis to extract the motif for each bin. This step is crucial for identifying the occurrences of bases at each position across the bins. Next, the consensus approach [14] 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.

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.

To address this, two possible approaches were derived, which are detailed below:

5.2 Method 2

The first approach was to analyse the 8mer PWMs generated. To do this, an understanding of which bases are present throughout the dataset and at what position is needed. First, weights have to be assigned to the bins to ensure that there is an expression of PWMs throughout the dataset and that the best PWM does not dominate it.

Here, the issue arises of how to choose the weights. After much deliberation, a metric is chosen to calculate the metric based on the E-Score values of the bins. Each bin contains a certain percentage of the dataset, and the E-score will vary unevenly in each bin. 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 with 4 bars denoting the four bases (A, C, G, T) for each of the eight positions. In this manner, the bases with the highest probability are the possible anchor residues.

5.3 Method 3: BINGO Algorithm

The other approach was to align the dataset alongside the top motif. A Python script was developed to go through the FASTA format generated for the top-enrichment 8mers file of the TF. It takes the top 8mer (that is, it has the highest binding factor) and generates substrings. Then, the substrings are ordered in a lexical manner and as per the length of the strings. Then, the subsequent sequences are aligned based on the longest possible substring alongside the top 8mer. This method recognises the sequence portion that occurs the most throughout the dataset. Therefore, the anchor residue and its position concerning the top 8mer is acquired.

The following is the algorithm for the approach discussed above:

Algorithm 1 Align Sequences to Motif

```
function ReadSequencesFromFile(filePath)
    sequences ← []
    for line in open(filePath) do
        if line[0] ≠'>' 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 ← []
    length ← len(baseStr)
    for i ← 0 to length − 1 do
        for j ← i + 1 to length do
            substrings.append(baseStr[i : j])
        end for
    end for
    return sorted(set(substrings), key = len, reverse = True)
    function AlignSequences(sequences, referenceSequence, motif)
        substrings ← GenerateSubstrings(motif)
        df ← DataFrame(columns=[−7 to 14])
        for sequence in sequences do
            sequence ← sequence.replace('.', '')
            aligned ← False
            entire ← False
            idx ← sequence.find(motif)
            if idx ≠ −1 then
                aligned ← True
                entire ← True
            else
                for substring in substrings do
                    if substring in sequence then
                        idx ← sequence.find(substring)
                        refIndex ← referenceSequence.find(substring)
                        aligned ← True
                        entire ← False
                        break
                    end if
                end for
            end if
            if aligned then
                offset ← refIndex − idx
                indices ← range(offset, offset + len(sequence))
                df.append(DataSeries(sequence, index=indices))
            end if
        end for
        return df
    end function
```

6 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

7 Results and Discussion

As GATA4 is a very well-known transcription factor, the PWM is very recognisable.

Figure 4: Protein Binding Microarray (PBM)-Derived DNA Binding Site Motif (Seed-And-Wobble) from UniProbe

In method 1, the discovered motifs were analysed using consensus sequence to characterise DNA-binding motifs across discrete 10

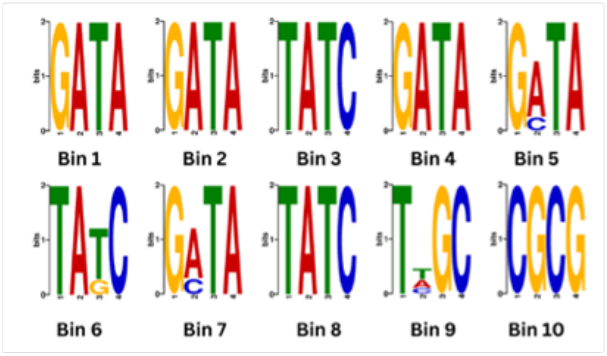

Figure 5: Motifs of Length 4 Identified in Each Discrete 10% Bin

For the 4mer motifs, consensus sequences ranged from "GATA" in the initial bins to a variation in the subsequent bins, including "TATC," "TTGC," and "CGCG," 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.

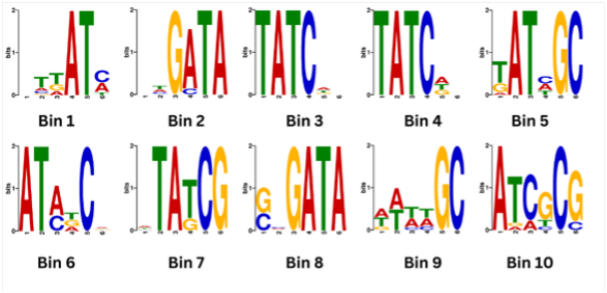

Figure 6: Motifs of Length 6 Identified in Each Discrete 10% Bins

Expanding the analysis to 6mer motifs, 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 first and second positions, respectively, suggesting a preference for these bases in the binding sites for transcription factors.

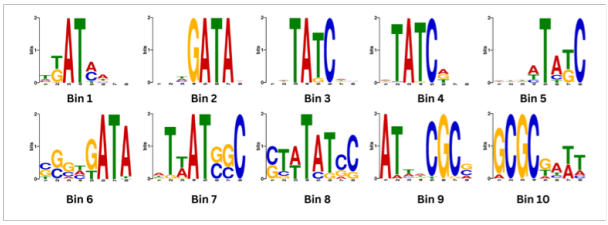

Figure 7: Motifs of Length 8 Identified in Each Discrete 10% Bin

Similarly, for 8mer motifs, 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.

So, it can be derived that the motif "GATA" or its reverse complement "TATC" are consistently identified across most bins, indicating accurate motif detection. However, the position of the anchor residues is not clear.

The graph below was generated using the weighted binning method.

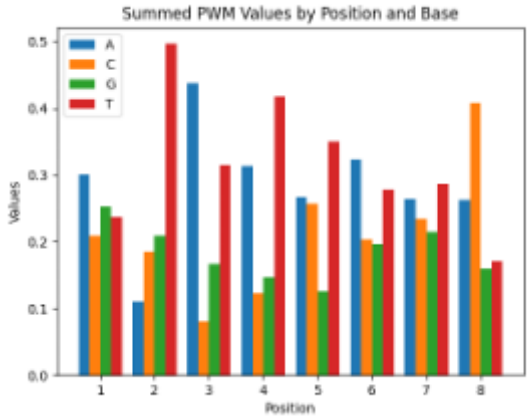

Figure 8: Summed PWM Values as per Weighted Bins

The logo below shows the possible anchor residue after aligning the GATA dataset.

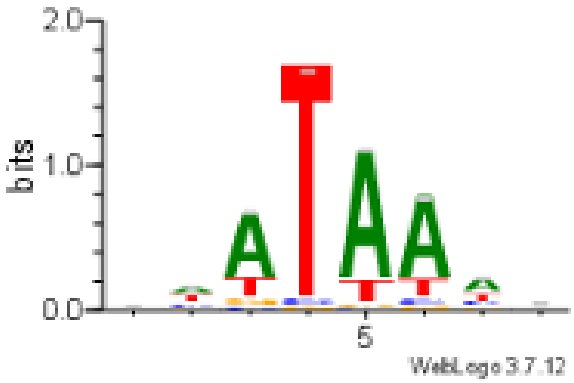

Figure 9: Sequence Logo after Aligning Entire Dataset

Based on the graph and sequence logo generated, we can conclude that 'T' at position 4 is the most prevalent across the entire dataset of GATA. This could be the anchor residue.

Understanding the placement of these anchors is crucial. This is where the method of aligning the sequences has helped. As per the sequence logo generated, it corroborates with the previous result of T being a possible anchor residue, and in addition, it can confirm the position.

8 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