Report
Submitted in fulfilment of the requirements of
CS F377 Design Project

By

Riddhi Goswami
IDNO: 2021A7PS0017U

Under the supervision of
Dr. J ANGEL ARUL JOTHI
Assistant Professor

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

DUBAI CAMPUS, DUBAI UAE

May - 2024

ACKNOWLEDGMENTS

First and foremost, I want to express my deepest gratitude to my Supervisor, Dr J Angel Arul Jothi, Assistant Professor in the Computer Science Department at BITS Pilani, Dubai Campus, United Arab Emirates, for her valuable guidance and encouragement during this project. I am extremely grateful to her for her able guidance, valuable technical input, and useful suggestions.

I express my sincere thanks and gratitude to our Director, BITS Pilani, Dubai Campus, Prof. Dr. Srinivas Madapusi, Dean ARD, and Dr. B. Vijaykumar, HOD, CS, for their motivation, encouragement, and support in pursuing my project.

Above all, I thank the Lord for giving me the strength to carry out this work to the best of my abilities.

Name: Riddhi Goswami
ID No. : 2021A7PS0017U

LIST OF FIGURES

Abstract

This project aims to advance the field of drug design by identifying and analysing anchor residues within transcription factor binding sites (TFBS) using a comprehensive examination of UniProbe data and the application of MEME Suite tools. The ultimate goal is to enable the development of novel therapeutic strategies that inhibit gene expression in diseased states by targeting these critical residues. Given the constraints of data availability, quality, and computational resources, this study employs a methodical approach to data procurement, preparation, and analysis. Initial steps involve collecting and pre-processing relevant data from UniProbe, focusing primarily on the human transcription factor GATA4. Subsequently, the project delineates the process of motif identification using MEME Suite's command-line interface, interpreting generated Probability Weight Matrices (PWM), and applying a consensus approach. The project introduces a novel binning strategy to discern anchor residues by comparing motif prevalence across various data segments, thus enabling the identification of pivotal anchor residues. This methodology sheds light on the intricate mechanisms of TFBS. It lays the groundwork for the potential design of drugs aimed at modulating gene expression through precise targeting of anchor residues. The deliverables of this project include a detailed report on the methodology, findings, and implications of these findings, complemented by visual representations of the data and adaptable code for broader dataset application. Through this endeavour, the project aspires to contribute significantly to the burgeoning field of targeted drug design, offering a promising avenue for developing treatments for various genetic disorders.

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.

I. 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. 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 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 marked substantial progress in predicting protein structures, yet translating these structures into predictable TF binding dynamics encompasses a myriad of complexities. Against this backdrop, our research introduces a novel methodology to predict anchor residues within TF binding sites, leveraging identified motifs from the UniProbe database. 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.

II. LITERATURE REVIEW

The field of computational biology has witnessed transformative advances, especially in the prediction and understanding of complex biological structures. These advancements are pivotal for decoding the myriad interactions that underpin biological functions and therapeutic interventions. This literature review delves into recent seminal works that epitomise the progress in this arena, highlighting methodologies, achievements, and their broader implications.

Accurate Prediction of Protein-Nucleic Acid Complexes Using RoseTTAFoldNA [6]

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. Building upon the foundational RoseTTAFold method, this technique represents a significant stride towards understanding complex biological structures, offering improved accuracy over prior models, including AlphaFold. RoseTTAFoldNA employs a sophisticated deep-learning architecture.

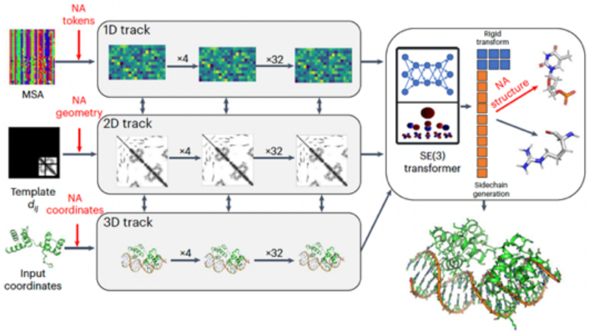


Fig. 1 | Overview of the architecture of RoseTTAFoldNA

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, culminating in a model with 67 million parameters.

This comprehensive approach enables the accurate modelling of intricate protein-nucleic acid interactions. The model performs better in predicting protein-nucleic acid complex structures, surpassing other state-of-the-art methods. It offers a novel benchmark in computational biology, emphasising the potential of deep learning in understanding biological complexities.

Highly Accurate Protein Structure Prediction with AlphaFold [7]

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 enhances our understanding of protein folding mechanisms and opens new avenues for biological research 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. Its training regime was multifaceted, 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).

Additionally, 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, offering an extensive resource of 65,983,866 families represented through MSAs and hidden Markov models, covering over 2.2 billion protein sequences from a variety of sources including reference databases, metagenomes, and metatranscriptomes. The training also incorporated an MSA Depth Analysis utilising the per-residue normalised number of effective sequences (Neff) for enhancing prediction accuracy.

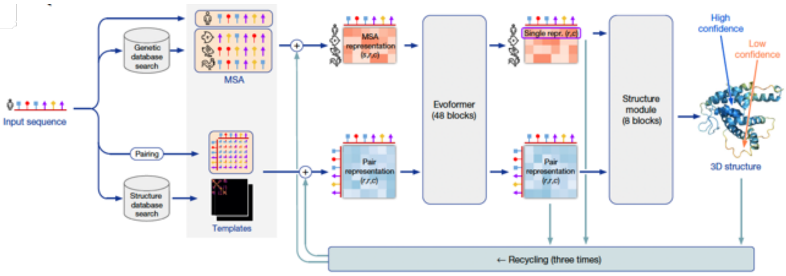

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

Fig. 2 | Model architecture of AlphaFold

AlphaFold's innovative architecture and methodologies mark significant departures from traditional models, 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. It 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 its profound impact on computational biology.

AlphaFold's accuracy and efficiency in protein structure prediction stand as a testament to AI's potential in revolutionising computational biology. Its ability to predict near-experimental accuracy structures paves the way for groundbreaking research in understanding biological processes and drug discovery.

Identification of Human Lineage-Specific Transcriptional Coregulators [8]

This section reviews the paper on identifying human lineage-specific transcriptional coregulators, highlighting the development of GENRE and a glossary of TF-8mer modules. These tools enhance the precision of identifying transcription factor binding sites, offering new insights into gene regulation and transcriptional coregulators. The paper's contributions lie in its novel approach to constructing genomic backgrounds and identifying TF binding motifs. The research advances our understanding of gene regulatory mechanisms by providing tools like GENRE, suggesting a methodology for studying tissue-specific gene-regulatory programs and potential therapeutic targets.

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

III. DATASET DESCRIPTION

The UniPROBE database [9], Universal PBM Resource for Oligonucleotide Binding Evaluation, is a dedicated platform for collating and sharing comprehensive data on protein-DNA interactions. Central to its repository is information derived from universal protein-binding microarray (PBM) experiments, which elucidate the binding affinities of proteins—primarily transcription factors (TFs)—across all conceivable DNA sequences of a given length, known as k-mers. This data is vital for deciphering gene regulation mechanisms, given the pivotal role of transcription factors in modulating gene expression.

The dataset encompasses a variety of files integral to understanding transcription factor-DNA interactions.

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. [10]

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:

1) `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. Here is what each column represents:
1. 8-mer: This column lists the 8-nucleotide sequences that were tested for binding affinity with the GATA4 protein

2. 8-mer (Complement): reverse complement, with A pairing with T and C pairing with G, read in the opposite direction.
3. E-score: The enrichment score quantifies GATA4's 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.
4. 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 GATA4 and the DNA sequence.
5. 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.

2) `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. This information is crucial for understanding the specific DNA sequences GATA4 targets, which can inform studies on gene regulation, transcription factor networks, and the molecular mechanisms underlying GATA4's roles in development and disease.

3) `GATA4_anti-GST_alldata.txt` file:

This file includes detailed experimental data for each microarray spot, including control and experimental probes.
1. Column & Row: These columns likely indicate the physical location of the probe on the microarray chip, helping to identify and correlate specific spots with their experimental data.
2. Name & ID: These fields provide identifiers for each probe or control spot on the array. "GE_BrightCorner" and "DarkCorner" are control spots. (The critical control point is transcription, the process of copying a gene's DNA sequence (transcribed) into an RNA molecule.)
3. Sequence: This column lists the DNA sequence associated with each probe. For experimental probes, this sequence is what the GATA4 protein would potentially bind to.
To understand the following attributes, we must understand fluorescence intensity measurements [11]. Fluorescence intensity measurements refer to quantifying the light emitted by certain molecules (fluorophores) when they are excited by light of a specific

wavelength. They are a standard method to detect and quantify various substances, including proteins, nucleic acids, and small molecules. Cy3 and Alexa488 are examples of fluorophores.
4. Cy3 & Alexa488: These columns show the fluorescence intensity measurements for each probe, detected by different fluorophores (Cy3 and Alexa488).
5. Cy3Flags & Alexa488Flags: Flags indicating the quality or reliability of the fluorescence measurements. A value of 0 typically means that the data point is considered reliable.
6. Cy3Exp: This could refer to expected values or an experimental condition related to Cy3 fluorescence.
7. Obs/Exp: It compares the fluorescence intensity to an expected value or model prediction. This could help in assessing the relative binding affinity of GATA4 to the probe sequences.
8. Alexa488Norm: Normalised Alexa488 fluorescence intensity adjusts the raw data to account for experimental variations, making comparisons between probes more reliable.
9. Alexa488Median: This might represent the median value of Alexa488 fluorescence across multiple measurements or experiments, providing a robust central tendency measure of the data.
10. Alexa488Adjusted: The Alexa488 fluorescence intensity was adjusted, possibly correcting the data for background noise or other experimental factors to accurately reflect the binding affinity.
The detailed data, including fluorescence intensities, normalisation, and adjustments, are crucial for understanding the specificity and dynamics of GATA4's DNA-binding activities.

4) `GATA4_anti-GST_combinatorial.txt` file:

It represents the results of combinatorial analysis from a protein binding microarray (PBM) experiment involving the GATA4 transcription factor.
Here is an interpretation of the format: The first Column represents a measure of binding affinity, fluorescence intensity, enrichment scores, or another metric indicating how strongly GATA4 binds to the corresponding DNA sequence. The second column has the DNA sequences tested in the experiment. Each sequence is unique and has been analysed to determine its binding affinity to GATA4. The standard part at the end of each sequence, `TCTGTGTTCCGTTGTCCGTGCTG`, might be a fixed region used in the experiment, possibly a scaffold or a part of the sequence necessary for the binding assay. In contrast, the varying parts at the beginning of each sequence represent the combinatorial variations being tested for GATA4 binding affinity. It highlights specific sequences GATA4 binds with varying degrees of affinity. It focuses on how changes in the DNA sequence (potentially through a combinatorial approach where different nucleotides are systematically varied) affect GATA4's binding ability.

5) `GATA4_anti-GST_RC.pwm` file:

This file gives the Position Weight Matrix (PWM) for the transcription factor GATA4, based on a specific seed k-mer (in this case, CCTTATCT) with a high enrichment score. PWMs are a

common bioinformatic representation used to describe the binding specificity of DNA-binding proteins, such as transcription factors.
1) Seed k-mer: CCTTATCT: DNA sequence that was used as a reference or starting point for the analysis
2) Enrichment Score: This score quantifies the binding affinity of the seed k-mer to GATA4, indicating that it is a strong binding site.
Following the header, the file lists the PWM, which consists of rows representing each of the four nucleotides (A, C, G, T) and columns representing each position within the binding site sequence. The values in the matrix give the frequency or probability of each nucleotide occurring at each position in sequences that bind to GATA4. The PWM is a powerful tool for predicting potential binding sites in DNA sequences by calculating how closely they match the preferred binding profile. This matrix is used to score sequences by calculating the log-likelihood of GATA4 binding to them, helping to identify potential GATA4 target genes and understand their role in gene regulation. The specificities captured in the PWM can inform on the biological functions of GATA4, including its involvement in heart development, cell differentiation, and other processes where GATA4 is a critical regulatory factor.
To reiterate, the binding affinity is the dependent variable, and the presence of 4-mer sequences is the independent variable.

6) `GATA4_anti-GST_regression.txt` file:

It contains the output from a regression analysis that models the relationship between different 4-mer DNA sequences (AAAA, AAAC, AAAG, etc.) and their binding affinities to the GATA4 transcription factor, as indicated by the coefficients (Theta values) associated with each sequence. The following are the interpretations of what theta entails [12].
- Theta[0='intercept']= 238.156284246807: The intercept (Theta[0]) represents the baseline value of the binding affinity when all the 4-mer sequences are at their reference level (typically zero).
- Theta[n='....']= Y: The file lists a Theta value representing the regression coefficient associated with each 4-mer sequence.
Positive Theta values indicate that the presence of the corresponding 4-mer sequence is associated with an increase in the binding affinity, for example, a higher binding affinity to GATA4.
Negative Theta values suggest that the presence of the 4-mer sequence is associated with a decrease in the binding affinity. This might indicate sequences to which GATA4 is less likely to bind.

IV. METHODOLOGY

The MEME Suite [13] stands as 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, which sheds light on the DNA-binding motifs critical for gene regulation. Additionally, it plays a crucial role in predicting protein functions by uncovering conserved motifs that hint at functional domains or active sites within protein sequences. Through comparative genomics and evolutionary studies, MEME Suite aids in tracing motif conservation across species, offering insights into evolutionary relationships and functional continuity. Furthermore, it assists in identifying regulatory elements within genomes, pinpointing regulatory motifs located upstream of genes to discern patterns of shared regulation.

This highlights the importance of the MEME Suite in advancing our comprehension of complex biological systems.
Major components of this suite include:
1) Multiple EM for Motif Elicitation (MEME): This tool is employed for identifying statistically significant, repeated patterns (motifs) within a set of sequences, and it helps discover novel motifs in related sequences.
2) TOMTOM: This feature compares discovered motifs against a database of known motifs to find matches and assess their statistical significance, aiding in the identification of motifs that resemble known binding preferences.
I have 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, I have to use a theoretical framework to identify the anchor residue.
Possible strategies include:
1) Consensus Approach [14]: Involves identifying the most prevalent residue at each position within a PWM and using consensus sequences to highlight critical binding affinities.
2) Information Content (IC) [15]: This is a measure of the specificity of residue occurrence at each position, with high IC values indicating potential anchor points.
3) Kullback-Leibler Divergence (KLD) [16]: Assesses the divergence of the observed residue distribution from an expected background distribution, with significant divergences pointing to anchor residues.

V. IMPLEMENTATION

For my implementation, I adopted a structured approach.

I used the "GATA4_anti-GST_8mers_top_enrichment.txt" file from the UniProbe dataset and converted it into FASTA format, which MEME requires for processing. I chose this file because 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 10% intervals.

I worked on a Python script to use the official MEME Docker container to help 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. I considered a range of 4,6,8 as the possible motif widths of interest. The number of motifs to be identified by MEME is another crucial input. I recommend starting with a smaller number for initial explorations to focus on the most significant motifs (I chose a value of 1 to get the most significant 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 command line, a sample command is:

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

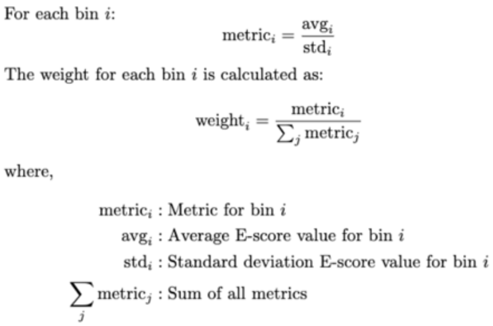
This instruction tells MEME to seek only one motif within the DNA sequences, with motif widths spanning eight nucleotides, within a dataset optimised for analysis efficiency.

The file meme.html is the most significant among the generated outputs as it encapsulates the motif discovery results. I created a Python script 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. Then, I can identify bases that consistently appear through the bin – these could most probably be the anchor residues.

An issue faced with this implementation course is that the position of the anchor residue cannot be determined yet. For this, I worked on two possible approaches, detailed below.

The *first approach* was to try to analyse the 8mer PWMs generated. To analyse this, we need to understand which bases are present throughout the dataset, and in which position. To this, we need to first assign weights to the bins. This is to ensure that there is expression of pwms throughout the dataset – and that it is not dominated by the best pwm.

Here, we face the issue of how to assign weights. After much deliberation, I chose to calculate the metric based on the E-Score values of the bins. Bins are created to ensure a certain percentage of the dataset is present in each bin. Thus, we need to understand how the E-score are values varying with each bin. For this reason, I decided to base the weights on the average and standard deviation of the E-Score values of the sequences present.



For each bin i:

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

The weight for each bin i is calculated as:

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

where,

metric_i : Metric for bin i

avg_i : Average E-score value for bin i

std_i : Standard deviation E-score value for bin i

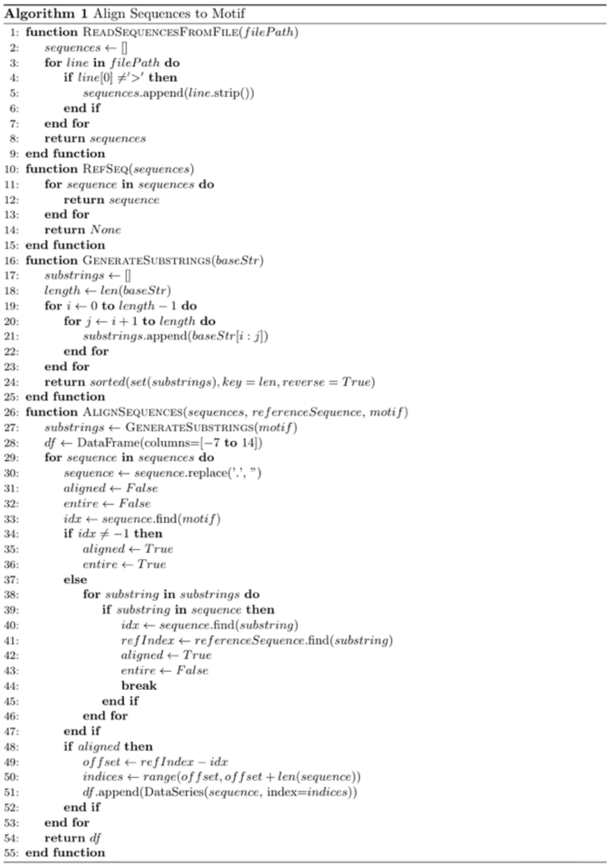
$\sum_j \text{metric}_j$: Sum of all metrics

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 8 positions. In this manner, we can concur that the bases with highest probability as the possible anchor residue.

The *second approach* was to align the dataset alongside the top motif. I worked on a python script 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, generates substrings of it. Then, I order the substrings in a lexical manner, and as per length of strings. Then, I will align the subsequent sequences based on the longest substring alongside the top 8mer.

Through this method, we end up with the portion of the sequence which occurs throughout the dataset. Through this method, we will come to know the anchor residue, as well the position of it, with respect to the top 8mer.

Given alongside is the algorithm:



Algorithm 1 Align Sequences to Motif

```
function READSEQUENCESFROMFILE($filePath$)
    $sequences \leftarrow []$
    for $line$ in $filePath$ 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
```

VI. EVALUATION METRICS

As I have used a well-known Transcription Factor GATA4, it is already known what the PWM should look like.

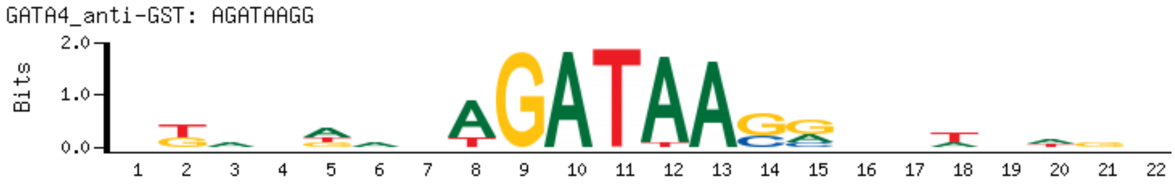


Fig. 3 | Protein Binding Microarray (PBM)-Derived DNA Binding Site Motif (Seed-And-Wobble) from UniProbe

The below PWMs represent the results procured through MEME for discrete bins of 10%. The change in motif length has caused the differences in PWMs. The length is mentioned alongside the figures.

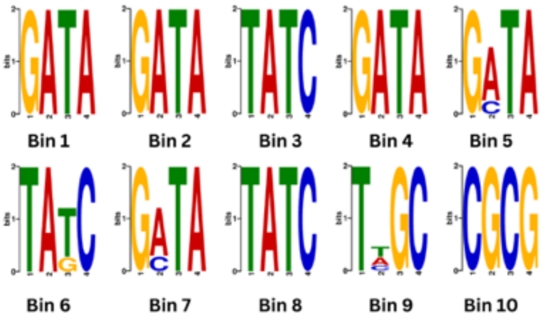


Fig. 4 | Motifs of Length 4 Identified in Each Discrete 10% Bin

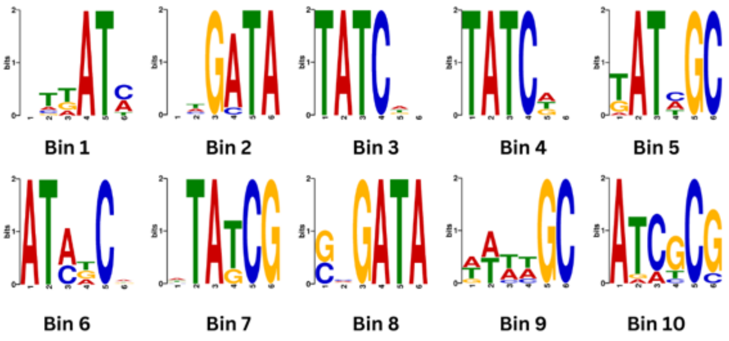


Fig. 5 | Motifs of Length 6 Identified in Each Discrete 10% Bin

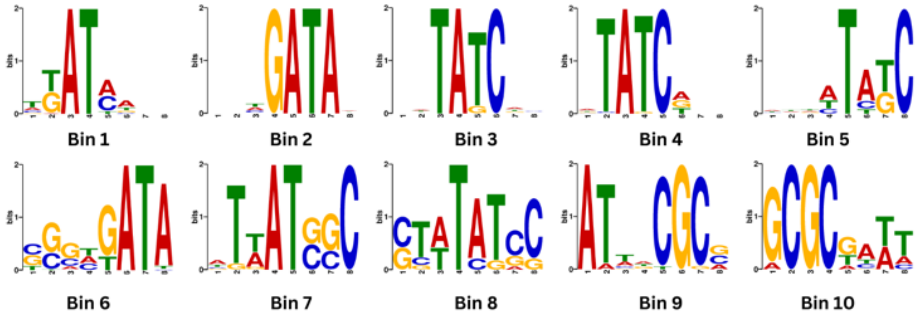


Fig. 6 | Motifs of Length 8 Identified in Each Discrete 10% Bin

The figures' analysis reveals that the motifs GATA or its reverse complement TATC are consistently identified across most of the bins, indicating accurate motif detection.

The figures' analysis reveals that the motifs GATA or its reverse complement TATC are consistently identified across most bins, indicating accurate motif detection. However, we cannot be sure of the positions of the anchor residues. To combat this, I derived the results through the two approaches discussed earlier.

The below graph was generated as a result of the weighted binning method.

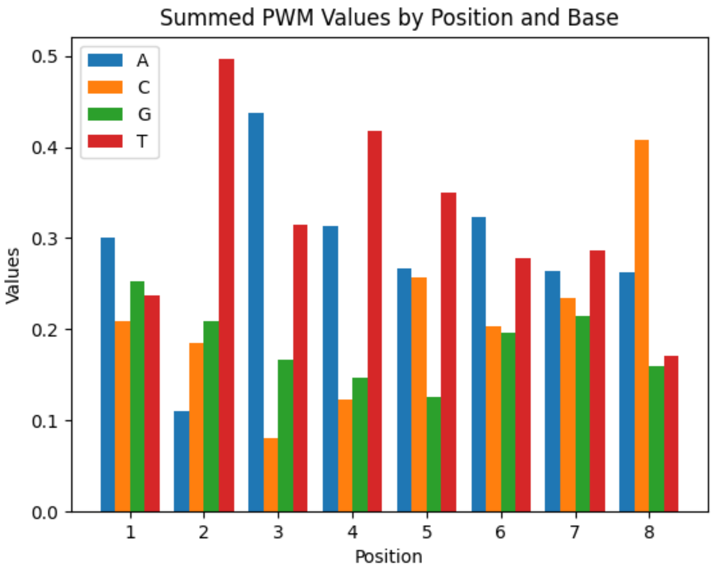


Fig. 8 | Summed PWM Values as per Weighted Bins

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

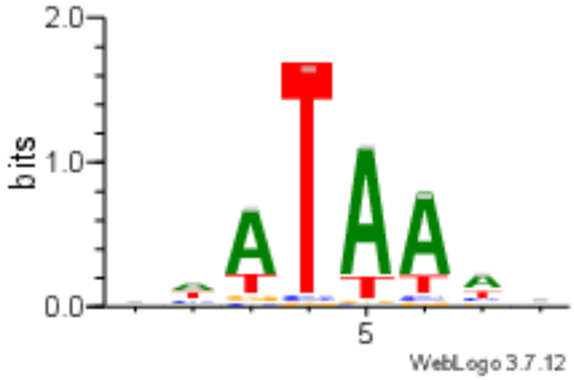


Fig. 9 | Sequence Logo after Aligning Entire Dataset

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.

VII. RESULTS AND DISCUSSIONS

In analysing and discussing previously discovered motifs with lengths of 4, 6, and 8 nucleotides, I employed consensus sequence analysis to characterise DNA-binding motifs across discrete 10% bins of the entire dataset.

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.

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.

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.

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 of result of T being a possible anchor residue, and in addition, it can confirm the position.

VIII. 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 as per the top motif, as presented in the project. This methodology was applied across the entire ETS domain of proteins, and the results are being 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][18][19] or other predictive methods could offer a sophisticated avenue for determining the locations of these anchor residues within the motifs

IX. 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. O., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., ... 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