Report
Submitted in fulfilment of the requirements of
CS F366 Laboratory 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

October - 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

Abstract

This project aims to identify and analyze anchor segments within transcription factor binding sites (TFBS) by leveraging both in vitro and in vivo data from UniProbe and CIS-BP databases. These datasets include data from protein-binding microarrays (PBM), systematic evolution of ligands by exponential enrichment (SELEX), and chromatin immunoprecipitation sequencing (ChIP-Seq), facilitating a comprehensive understanding of monomeric and dimeric transcription factors (TFs) and their binding patterns on both gapped and ungapped motifs. Building upon a novel alignment algorithm, this study enhances the algorithm to accurately account for dimer binding in both continuous and gapped forms, refining predictions of binding affinity and specificity. Our primary objective is to advance drug design by identifying essential anchor segments critical to gene expression, thereby enabling the development of therapeutics that restrict TF binding. Additionally, we will deploy a Django-based web server that allows users to input TF data, utilize the anchor residue finder, and visualize 3D models highlighting TF-DNA interactions. This project contributes a valuable tool for targeting TF binding sites, with implications for addressing disorders associated with transcriptional gene regulation.

Keywords

Transcription Factor Binding Sites (TFBS); Anchor Residues; UniProbe Database; 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

Transcription factors (TFs) are essential proteins that play a pivotal role in regulating gene expression by binding to specific DNA sequences known as transcription factor binding sites (TFBS). These interactions are fundamental to various cellular processes, including development, immune responses, and cellular differentiation, and are increasingly recognized as therapeutic targets due to their involvement in many diseases, including cancer, neurodegenerative disorders, and immune dysregulation [1]. Recent advancements in structural biology and genomics have significantly enhanced our understanding of TF-DNA interactions, particularly through the availability of high-resolution data obtained from technologies such as protein-binding microarrays (PBM), systematic evolution of ligands by exponential enrichment (SELEX), and chromatin immunoprecipitation sequencing (ChIP-Seq). These methods, utilized in databases like UniProbe and CIS-BP, provide both in vitro and in vivo insights into TF binding affinities, sequence preferences, and structural specificity, enabling comprehensive analyses of monomeric and dimeric TF interactions across various sequence motifs [2] [3]. Understanding the specificity and affinity of TF binding is further refined by examining anchor segments—regions within TFBS that are critical for stable TF-DNA interactions. These anchor segments are hypothesized to contribute significantly to the selectivity of TF binding by either directly engaging DNA or stabilizing the protein-DNA complex through allosteric interactions. Current research is increasingly focusing on how variations in these anchor regions can influence binding affinity and specificity, particularly in complex motifs with gapped and ungapped patterns. Novel alignment algorithms are now being developed to identify and characterize these critical anchor residues, facilitating predictions of TF binding in both continuous and discontinuous DNA motifs [4].

This study builds on recent advances by refining our alignment algorithm to better model) TF-DNA interactions, with a particular focus on accurately capturing dimer binding. By identifying key anchor segments, we aim to support the development of therapeutics that can selectively inhibit TF binding, potentially enabling new approaches to modulate gene expression in transcription-related disorders. To enhance accessibility, we will develop a Django-based web server where researchers can input TF data, identify anchor residues, and visualize 3D TF-DNA interactions. This platform will serve as a valuable resource for exploring and potentially targeting TF binding sites for therapeutic applications.

The various sections in report explain the reason behind this project and how the problem statement is tackled. Section I gives an overview of the significance of TFBS in gene regulation. Section II outlines the recent advances revealing the many factors that affect choice of TFBS. Section III describes the UniProbe dataset chosen as input for the algorithm. Section IV explains the algorithm in detail and how various improvements are included for robustness and better accuracy.

II. LITERATURE REVIEW

TFs bind to DNA at specific sites to regulate gene expression, a process influenced by both the sequence of DNA and its structural context. While early research primarily focused on identifying core motifs (short sequences essential for TF recognition) recent studies emphasize the importance of sequences flanking these motifs and DNA flexibility in modulating TF binding specificity. This review examines the major findings across recent studies, providing insight into how TF-DNA interactions are determined by factors extending beyond core binding motifs.

Influence of Sequence Context on TF Binding: The specificity of TF binding is significantly affected by sequence context, particularly by flanking regions adjacent to core motifs. Gordân et al. (2013) [5] investigated the DNA binding behavior of basic helix-loop-helix (bHLH) TFs and found that flanking sequences alter DNA shape, thus influencing binding affinity. The study demonstrated that these non-core regions contribute to the shape of the DNA at binding sites, highlighting the role of three-dimensional DNA structure in TF specificity. Similarly, Santolini et al. (2014) [6] proposed a pairwise interaction model to capture nucleotide correlations within flanking sequences, which improved predictions of TF binding sites by accounting for interactions between nucleotides in these regions. Additionally, studies such as Hüntelmann et al. (2014) [7] have examined the role of conserved residues within DNA-binding motifs that facilitate stable binding across TF families. In their work, conserved lysine residues in the STAT1 TF were identified as crucial for DNA recognition, emphasizing that sequence conservation within motifs is as important as the surrounding context.

Role of Intrinsically Disordered Regions in TF Specificity: TF specificity can also be influenced by domains outside the core DNA-binding region. Brodsky et al. (2023) [8] explored how intrinsically disordered regions (IDRs) beyond the DNA-binding domain contribute to TF specificity. These regions can adopt multiple conformations, allowing for flexible interactions with diverse DNA sequences and adapting binding affinity in response to cellular conditions. This study underscores the need for predictive models that incorporate non-canonical domains in TFs, as these regions may contribute to subtle changes in binding specificity that core domains alone do not explain.

Enrichment of TF Binding Sites in Non-Coding Regions: Expanding on the genomic context of TF binding, Piriyapongsa et al. (2011) [9] revealed that TF binding sites are highly enriched within microRNA precursor sequences. This discovery suggests a regulatory interplay between TFs and non-coding RNAs, such as microRNAs, which are involved in post-transcriptional regulation. This connection extends the function of TFs beyond traditional gene expression and positions them as significant players in broader regulatory networks.

Advances in Computational Modeling of TF Binding Sites: Predicting TF binding sites has traditionally relied on Position Weight Matrices (PWMs), which score binding probabilities for each nucleotide position in a motif. However, PWMs assume nucleotide independence, oversimplifying complex TF-DNA interactions. Recent advances in computational modelling address these limitations by considering nucleotide interactions and incorporating flanking sequences and DNA structure. Santolini et al. (2014) improved predictive models with their pairwise interaction model, which captures dependencies between adjacent nucleotides. Machine learning approaches, including deep learning methods like convolutional and recurrent neural networks, have further enhanced TF binding predictions. These models, exemplified by He et al. (2024) [10] with their CRA-KAN hybrid model, incorporate attention mechanisms to identify key regions, refining predictions of TF binding specificity. DNA shape analysis has also become integral to TF binding prediction, as it accounts for the three-dimensional conformation of DNA, an influential factor in binding affinity. Hidden Markov Models (HMMs) and other machine learning techniques now combine sequence motifs, DNA flexibility, and structural information to enable genome-wide predictions that align closely with in vivo binding observations.

In conclusion, the reviewed research underscores that TF binding specificity is influenced by a range of factors beyond core motifs, such as flanking sequences, DNA structural properties, and non-canonical TF domains. Building on these insights, our approach introduces the concept of anchor residues, hypothesizing that within the multiple motifs of a TF, certain sequences remain consistent and are essential for binding. Our alignment algorithm aims to identify these anchor residues and validate our findings through protein crystallography, potentially advancing our understanding of TF binding mechanisms.

III. DATASET DESCRIPTION

The UniPROBE database [11], 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 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 TFs in modulating gene expression.

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

This includes files containing the TF'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 TF and various DNA sequences. Enrichment scores within the dataset provide a comparative analysis of a sequence's binding affinity to the TF, 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 TF's novel binding preferences. [12]

Among the many datasets, the file GATA4_anti-GST_8mers_11111111.txt is of importance. The following are the columns in the file:

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 TF, 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.

IV. METHODOLOGY

A Python script was developed to facilitate the processing and visualization of TF binding site data, and subsequently identifies . It includes core functions designed to read, align, and visually represent sequences, followed by a main workflow that iteratively processes TFsto align sequences and generate visual sequence logos.

The function **read_sequences_from_file(file_path, col_index)** reads DNA sequences and their scores from a specified file. It accepts a file path as input and an orientation indicator (col_index), where 0 denotes a forward sequence and 1 indicates a reverse sequence. This function returns two lists: one with the DNA sequences extracted from the file and another with corresponding scores, which are used to rank and select the seed for alignment. To support the alignment process, **generate_sorted_substrings(s)** generates all unique substrings of a given DNA sequence, sorts them by descending length. This function takes a single sequence as input and returns a list of sorted, unique substrings to facilitate optimal alignment with the seed. The **align_sequences(sequences, seed)** function aligns each sequence from the list of input sequences to the seed. After an alignment, the common number of bases (at the same position) in both seed and sequence is counted. Based on the highest number of common bases, that alignment is chosen for this sequence. These alignments are added to a Pandas DataFrame where each column represents positional offsets for alignment. This DataFrame serves as the basis for further visualization and analysis. Finally, the **sequence_logo_generator(TF, file_path, col_index)** function generates a sequence logo from the aligned sequences, saving the result as a PNG image. This function uses the TF to name the output file, along with a path to the file containing aligned sequences and a column index specifying the orientation of the sequences (forward or reverse).

In the main processing workflow, for each TF, it processes both orientations (regular and reverse complement) by setting the col_index to either 0 or 1. The workflow begins by reading and extracting sequences and corresponding E-scores from the input file. The seed is chosen based on the highest E-score. When col_index is set to 0, the seed is taken in its original form; if set to 1, seed is taken as its reverse. Against this seed all the other sequences are aligned using the align_sequences function, resulting in a DataFrame of aligned sequences. Once aligned, the sequences are formatted and saved to an output file for further analysis. The final step in the workflow generates a sequence logo from the aligned data. This is achieved using the sequence_logo_generator function, which visualizes the DNA-binding preferences of the TFs as a PNG file. Throughout the script, error handling ensures robustness by raising a **ValueError** if no seed is found during the alignment step.

The algorithm was enhanced through several significant modifications. First, an alignment check was introduced, focusing on the number of common characters (i.e. same bases with the same position) between the seed and the sequences to improve accuracy. Second, complement sequences were added in reverse, enabling the inclusion of dimer binding considerations. This modification aligns with the hypothesis that in dimer binding, the anchor residues of both the sequence and its complement should match, while for monomer binding, the anchors are positioned at opposite ends. Additionally, code was directly implemented for generating sequence logos, utilizing the WebLogo library [13] for streamlined pipeline.

VI. REFERENCES

[1] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. Cell, 172(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

[2] Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., & Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. The EMBO Journal, 29(13), 2147–2160. https://doi.org/10.1038/emboj.2010.106

[3] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, Kazuhiro R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, Juan M., Vincentelli, R., Luscombe, Nicholas M., Hughes, Timothy R., Lemaire, P., Ukkonen, E., Kivioja, T., & Taipale, J. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell, 152(1-2), 327–339. https://doi.org/10.1016/j.cell.2012.12.009

[4] Aditham, A. K., Markin, C. J., Mokhtari, D. A., DelRosso, N. V., & Fordyce, P. M. (2020). High-throughput binding affinity measurements for mutations spanning a transcription factor-DNA interface reveal affinity and specificity determinants. BioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2020.06.22.165571

[5] Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. 2013 Apr 25;3(4):1093-104. doi: 10.1016/j.celrep.2013.03.014.

[6] Santolini, M., Mora, T., & Hakim, V. (2014). A General Pairwise Interaction Model Provides an Accurate Description of In Vivo Transcription Factor Binding Sites. PLoS ONE, 9(6), e99015. https://doi.org/10.1371/journal.pone.0099015

[7] Hüntelmann, B., Staab, J., Christoph Herrmann-Lingen, & Meyer, T. (2014). A Conserved Motif in the Linker Domain of STAT1 Transcription Factor Is Required for Both Recognition and Release from High-Affinity DNA-Binding Sites. PLoS ONE, 9(5), e97633–e97633. https://doi.org/10.1371/journal.pone.0097633

[8] Brodsky, S., Jana, T., & Barkai, N. (2021). Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. Current Opinion in Structural Biology, 71, 110–115. https://doi.org/10.1016/j.sbi.2021.06.011

[9] Jittima Piriyapongsa, Jordan, I. K., Conley, A. B., Ronan, T., & Smalheiser, N. R. (2011). Transcription factor binding sites are highly enriched within microRNA precursor sequences. Biology Direct, 6(1). https://doi.org/10.1186/1745-6150-6-61

[10] He, G., Ye, J., Hao, H., & Chen, W. (2024). A KAN-based hybrid deep neural networks for accurate identification of transcription factor binding sites. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-4664531/v1

[11] 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

[12] 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

[13] WebLogo 3 - User’s Manual. (n.d.). Weblogo.threeplusone.com. https://weblogo.threeplusone.com/manual.html