# Sequence and pharmacophore analysis of DNA recognition helices in the HTH family of proteins and their binding DNA

**A PROJECT REPORT**

*Submitted by:*

**Suvan Kumar Sahu**

**(235HSBB034)**

*to*

**Institute of Bioinformatics and Applied Biotechnology**

*in partial fulfilment of the requirements for*

Master of Science in Biotechnology and Bioinformatics (Degree to be awarded by Bangalore University, Bengaluru)

*under the guidance of:*

**Prof. S Thiyagarajan**

IBAB
Institute of Bioinformatics and
Applied Biotechnology

**INSTITUTE OF BIOINFORMATICS AND APPLIED BIOTECHNOLOGY BENGALURU**

**May 2025**

# DECLARATION

I certify that

a) the work contained in this report is original and has been done by me under the guidance of my supervisor.

b) I have followed the guidelines provided by the Department in preparing the report.

c) I have conformed to the norms and guidelines given in the Honor Code of Conduct of the Institute.

d) whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

*Signed by:*

**Suvan Kumar Sahu**

# CERTIFICATE

It is certified that the work contained in this report titled “Sequence and pharmacophore analysis of DNA recognition helices in the HTH family of proteins and their binding DNA” is the original work done by Suvan Kumar Sahu and has been carried out under my supervision.

*Signed by:*

**Prof. S. Thiyagarajan**

**Date: 28.05.25**

# ABSTRACT

The helix-turn-helix (HTH) motif is a highly conserved structural domain found in a wide range of DNA-binding proteins across prokaryotes and eukaryotes. This project aims to investigate the sequence and pharmacophore patterns of HTH motifs, particularly within the recognition helices responsible for specific DNA interaction. We collected 155 HTH sequences, performed redundancy filtering, and extracted the DNA recognition helices using a custom Python pipeline. Each amino acid was categorized into one of nine pharmacophore classes based on physicochemical properties, and a corresponding pharmacophore sequence was generated for each motif. Using a quantitative matrix-based scoring system, we aligned and clustered the pharmacophore sequences to identify dominant patterns and structural conservation. Structural interaction analyses were conducted using data from the Protein Data Bank (PDB), focusing on side-chain interactions and atomic contacts at the DNA-protein interface. The resulting pharmacophore matrix and motif alignment scores reveal distinct clusters of HTH motifs with potential functional and evolutionary significance. This work provides a foundation for predictive modelling of DNA-binding specificity and offers a novel pharmacophore-centric approach to motif classification in transcriptional regulators.

# ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to **IBAB** and the **Government of Karnataka** for providing an exceptional platform for learning and skill development. The resources, mentorship, and academic environment offered have greatly enriched my research experience.

I am deeply thankful to **Prof. S. Thiyagarajan** for giving me the opportunity to be part of this project. His guidance, support, and encouragement played a key role in making this project possible.

I would also like to sincerely thank everyone who supported me throughout the course of this project, with special appreciation to **Mr. Labhendra Tiwari** and **Mr. Aman Vishwakarma** for their valuable assistance and cooperation.

# TABLE OF CONTENTS

Page

# List of figures

# List of tables

# List of Abbreviations

| **Abbreviation** | **Description** |
|---|---|
| API | Application Programming Interface |
| BLOSUM | Blocks substitution matrix |
| CAP | Catabolite gene activator protein |
| CD-HIT | Cluster Database at High Identity with Tolerance |
| CRP | Cyclic-AMP Receptor |
| CSV | Comma-separated values |
| cAMP | cyclic-AMP |
| DNA | Deoxy-ribonucleic acid |
| Fis | Factor for inversion stimulation |
| HTH | Helix-turn-helix |
| MEME | Multiple Expectation-Maximization for Motif Elicitation |
| NAKB | Nucleic Acid Knowledge Base |
| PDB | Protein Data Bank |
| RNA | Ribonucleic acid |

# Chapter 1
# Introduction

## 1.1 Helix-turn-helix motifs

DNA-binding motifs are structural domains present within proteins that enable them to interact with DNA, thereby playing a crucial role in the regulation of gene expression, replication, and DNA repair. These motifs have a unique ability to recognize and bind to specific DNA sequences through highly specific molecular contacts. One of the most common and well-characterized DNA-binding motifs is the helix-turn-helix (HTH), which is widely found in transcription factors and plays a significant role in gene regulation in both prokaryotic and eukaryotic organisms.

The HTH motif typically comprises two alpha helices that are connected by a short, flexible turn. The first helix, often referred to as the stabilizing helix, functions as a structural scaffold that helps anchor the protein-DNA complex, while the second helix, known as the recognition helix, makes sequence-specific contacts with the base pairs located in the major groove of the DNA. The stabilizing helix interacts with the minor groove of the DNA, thereby assisting in properly positioning the recognition helix so that it fits into the major groove, while the turn provides the necessary flexibility to maintain the overall structure of the HTH motif. The turn in an HTH domain is not a typical turn that reverses the direction of the polypeptide chain; rather, it is a short linker between the two α-helices that helps orient them correctly, particularly positioning the recognition helix for DNA binding.

The recognition helix carries out sequence-specific interactions with DNA bases, relying on hydrogen bonding, van der Waals forces, and electrostatic interactions. The sequence and structural variations within the recognition helix are key determinants of binding specificity to DNA motifs, which ultimately influences gene regulatory mechanisms. As a result, performing sequence analysis on the recognition helix can reveal conserved residues that are critical for both DNA binding and sequence specificity, offering important insights into protein-DNA interactions and gene regulation.

### 1.2 Oligomeric states of HTH proteins

Helix-turn-helix (HTH) proteins exhibit considerable diversity in their oligomeric states, which plays an important role in determining their DNA-binding specificity and regulatory capabilities. These oligomeric forms- monomers, homodimers, heterodimers, and higher-order multimers, affect how HTH proteins recognize and interact with DNA, thereby influencing gene expression across a wide range of organisms.

Monomeric HTH proteins are the simplest form, containing a single HTH motif per polypeptide chain. These proteins function independently and often recognize asymmetric DNA sequences. Among the most prevalent and functionally significant are homodimeric HTH proteins. These consist of two identical subunits, each contributing an HTH motif that engages one half of a palindromic DNA sequence. The dimerization is typically stabilized through non-covalent interactions such as hydrophobic forces and salt bridges, often involving auxiliary domains like coiled-coils. This symmetric interaction enables the protein to bind DNA with increased specificity and affinity. Classical examples of such proteins include the Lac repressor (LacI), which regulates lactose metabolism in E. coli (Lewis et al., 1996), and the catabolite activator protein (CAP), which activates gene expression in response to cAMP levels (Busby & Ebright, 1999). The Cro repressor from bacteriophage λ is another well-known homodimer that controls the lytic-lysogenic switch by binding to operator DNA.

Homodimeric HTH proteins are particularly important because they enable cooperative DNA recognition, allowing subtle environmental signals (e.g., changes in metabolite concentrations) to produce strong regulatory outcomes. This is often achieved through allosteric regulation, where the binding of a ligand to one subunit can influence the DNA-binding activity of the entire dimer. Such mechanisms are crucial for enabling bacteria and other organisms to respond swiftly to metabolic changes or stress conditions. In addition, the symmetric nature of homodimers allows for the formation of higher-order structures like tetramers, which can induce DNA looping. This capability enables more complex gene regulation, such as repression over long genomic distances or coordinated control of multiple operons.

In contrast, heterodimeric HTH proteins consist of two different subunits, each contributing unique DNA-binding features. This configuration allows for combinatorial control over gene regulation, increasing the range of DNA sequences that can be recognized and providing flexibility in responding to diverse cellular signals. These are more common in eukaryotic systems, where fine-tuned gene expression is often required.

### 1.3 Role of Pharmacophore modelling

Pharmacophore properties refer to the key structural and chemical features of a molecule (such as a ligand or protein) that are responsible for its biological activity by enabling it to interact with a specific target (e.g., a protein, enzyme, or receptor). These include properties like aromaticity, hydrogen bond donors, hydrogen bond acceptors, positive and negative charges, etc.

Investigating the sequence and pharmacophore characteristics of these proteins is critical to understanding their DNA-binding mechanisms and sequence specificity. Advances in computational biology have made it possible to systematically analyse HTH protein sequences, revealing conserved motifs, crucial residues, and structural patterns linked to their functional roles.

Furthermore, pharmacophore modelling, which identifies the spatial arrangement of chemical features required for molecular interactions, offers a powerful means to study how HTH proteins bind to DNA and how small molecule inhibitors could disrupt this interaction. Such insights are fundamental not only for deciphering protein-DNA interactions but also for developing therapeutic strategies for diseases associated with transcriptional dysregulation. By combining sequence analysis with pharmacophore modelling, this research seeks to provide a comprehensive understanding of the structural and functional determinants of DNA-binding HTH proteins, paving the way for future studies into their biological significance and drug development potential.

### 1.4 Objectives

- Perform Sequence analysis of helix-turn-helix proteins
- Identify the pharmacophore pattern in the HTH recognition helix
- Derive a new scoring matrix based on pharmacophore properties
- Identify the nucleotide sequence or pattern to which HTH binds

# Chapter 2
# Literature Survey

## 2.1 Previous knowledge

The helix-turn-helix (HTH) motif is a structural domain present in most DNA-binding proteins and is essential for the regulation of transcription. Since the discovery, many studies have been dedicated to understanding the sequence and structural characteristics of this motif, providing details about its mechanisms of recognizing DNA and its evolutionary role.

Initial investigations described the structural mechanism for DNA recognition using the HTH motif, focusing on the importance of the recognition helix, which typically inserts into the major groove of the DNA to form sequence-specific hydrogen bonds and van der Waals interactions with the nucleotide bases. It set the ground for what can affect protein-DNA binding specificity based on DNA-binding helices' sequence diversity. But then, sequence-based HTH motif detection was in its early stages, and there was an obvious need for computational methods that could detect these motifs with greater accuracy (Brennan 1992).

The helix-turn-helix motif was first discovered in lambda repressor and Cro proteins from lambda bacteriophage. It consists of two alpha helices that are separated by a short turn. The two alpha helices are nearly perpendicular to each other. The HTH motifs have an average length of 22 amino acids, with a minimum length of 20 and maximum length of 32. The DNA-binding domains containing the HTH motifs are much larger (25-150 residues) (Pellegrini-Callace et al. 2005).

Experiments have shown that the two α-helices project out from the surface of the protein, creating a convex unit that is inserted into the DNA groove, and the phosphate backbone and recognition helix interactions help in stabilizing the DNA-binding process (Steitz et al. 1982). Early structural characterization of bacteriophage λ lambda repressor and Cro proteins initially recognized the HTH motif as a universally conserved DNA-binding component (Anderson et al. 1981). Further studies have determined that the HTH structure is quite well-conserved but its sequence is very variable among protein families so it is challenging to identify by sequence-based approaches alone (McLaughlin et al. 2003).

Several HTH-containing proteins have been found to create structural distortions in DNA, such as smooth bends and sharp kinks, that can have functional importance for transcriptional control. DNA bending helps in strengthening of protein-protein interactions and to attract transcription machinery in proteins like CAP (catabolite gene activator protein) and Fis (factor for inversion stimulation) (Brennan 1992).

Conserved residues have been found by several studies to contribute to the structural stability and functionality of the HTH motif. Alanine (A) and Glycine (G) at positions 5 and 6 make contributions to structural stability, whereas Glycine (G) at position 10 is essential for turn flexibility (Rosinski et al. 1999). Leucine (L), Valine (V), and Isoleucine (I) at residue 16 stabilize the hydrophobic core, while Aspartic Acid (D) or Glutamic Acid (E) at residue 4 provide electrostatic interactions. These conserved interactions are crucial in ensuring correct DNA binding (Rosinski et al. 1999).

Even though the HTH motif is mainly related to DNA-binding proteins, research has also indicated that some RNA-binding proteins have HTH-like folds, suggesting a wider functional application outside of DNA binding. HTH motifs are present in a number of transcription factors such as AraC, ArsR, AsnC, CRP, DeoR, GntR, IclR, LacI, LuxR, LysR, MarR, MerR, TetR, and sigma factors (Harrison et al. 1991).

Due to the variability in the sequence of HTH motifs, most research indicates that structure-based methods are more accurate than sequence-based methods to identify HTH-containing proteins (Pellegrini-Callace et al. 2005). Through the combination of structure and sequence data, computational models have greatly enhanced the identification of HTH motifs and have enabled the identification of new DNA-binding proteins.

### 2.2 Research gaps

In spite of these developments, there are a number of research gaps. One of the gaps is that the sequence of the recognition helix in HTH motifs is highly variable among different HTH-containing proteins. Therefore, no exact pattern or consensus sequence has been identified that characterizes the recognition helices of all HTH proteins. Such variability complicates the prediction of DNA-binding specificity through sequence analysis alone, restricting the accuracy of computational models for motif detection.

Addition to this, the precise DNA motifs to which HTH proteins bind are not fully characterized, as binding specificity appears to be influenced not only by the recognition helix sequence but also by the context of the flanking residues, protein-rotein interactions, and chromatin environment.

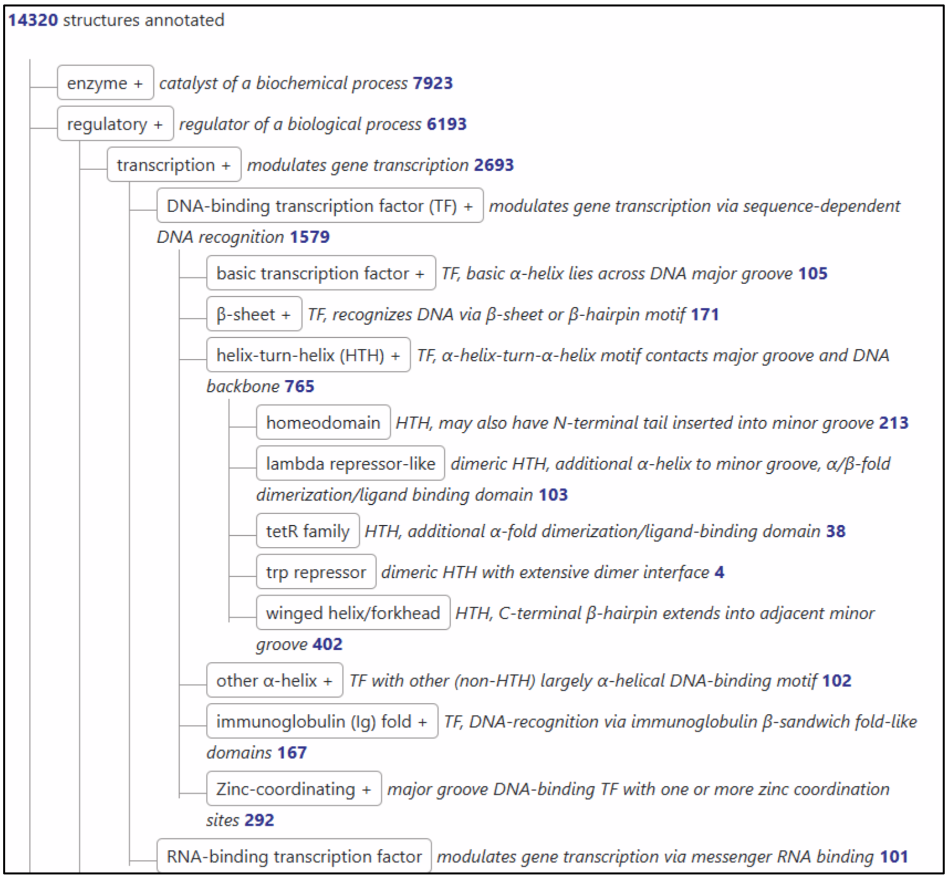
Another significant gap in the field is the lack of a complete understanding of the various modes of DNA binding adopted by different types of HTH proteins. While the classical model involves insertion of the recognition helix into the major groove of the DNA, structural studies have revealed alternative binding orientations and multi-protein assemblies that contribute to gene regulation, yet these mechanisms remain incompletely understood.

# Chapter 3

# Materials and Methods

## 3.1 Protein Sequence Collection

To initiate the analysis, protein sequences containing DNA-binding helix-turn-helix (HTH) motifs were retrieved from the Nucleic Acid Knowledge Base (NAKB), a curated database that provides detailed structural and functional annotation for nucleic acid-binding proteins. The NAKB database was queried to identify all regulatory proteins, among which transcription factors (TFs) and specifically DNA-binding transcription factors (DB-TFs) were filtered. From this refined set, only those DNA-binding transcription factors known to adopt helix-turn-helix (HTH) structural motifs were selected for further study.



**14320** structures annotated

- enzyme + *catalyst of a biochemical process* **7923**
- regulatory + *regulator of a biological process* **6193**
  - transcription + *modulates gene transcription* **2693**
    - DNA-binding transcription factor (TF) + *modulates gene transcription via sequence-dependent DNA recognition* **1579**
      - basic transcription factor + *TF, basic α-helix lies across DNA major groove* **105**
      - β-sheet + *TF, recognizes DNA via β-sheet or β-hairpin motif* **171**
      - helix-turn-helix (HTH) + *TF, α-helix-turn-α-helix motif contacts major groove and DNA backbone* **765**
        - homeodomain *HTH, may also have N-terminal tail inserted into minor groove* **213**
        - lambda repressor-like *dimeric HTH, additional α-helix to minor groove, α/β-fold dimerization/ligand binding domain* **103**
        - tetR family *HTH, additional α-fold dimerization/ligand-binding domain* **38**
        - trp repressor *dimeric HTH with extensive dimer interface* **4**
        - winged helix/forkhead *HTH, C-terminal β-hairpin extends into adjacent minor groove* **402**
      - other α-helix + *TF with other (non-HTH) largely α-helical DNA-binding motif* **102**
      - immunoglobulin (Ig) fold + *TF, DNA-recognition via immunoglobulin β-sandwich fold-like domains* **167**
      - Zinc-coordinating + *major groove DNA-binding TF with one or more zinc coordination sites* **292**
    - RNA-binding transcription factor *modulates gene transcription via messenger RNA binding* **101**

**Fig.3.1** Protein Annotation tree in NAKB

These HTH-containing proteins were then classified into five major structural subtypes based on their three-dimensional configurations: homeodomain, lambda repressor-like, TetR family, trp repressor, and winged helix/fork-head. The PDB IDs for each of these HTH proteins were compiled into a CSV file.

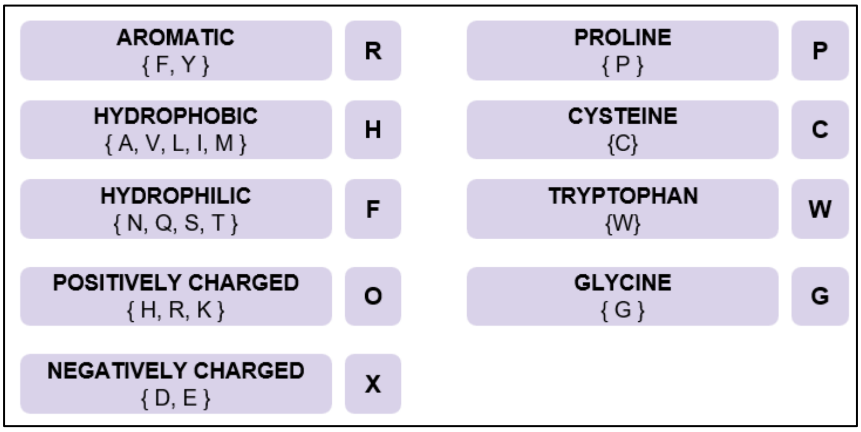
A python script was used to automate the retrieval of protein sequences. This script read the CSV file containing the list of PDB-IDs, accessed the corresponding PDB entries, and extracted the amino acid sequences of the proteins from the PDB database. All retrieved sequences were then compiled and stored in a FASTA-formatted file to create a curated dataset for further analyses.

### 3.2 Separating the homodimer proteins

Among the downloaded proteins, those forming homodimers were specifically identified and separated for further analysis. This classification was achieved using a custom Python script that searched the PDB database using the list of previously obtained PDB IDs. The script accessed the structural summary pages for each protein and parsed the Global Stoichiometry section to detect the keyword "Homo 2-mer", which indicates that the protein functions as a homodimer in its biologically relevant quaternary structure. This automated approach enabled efficient and accurate identification of homo-dimeric proteins from the larger dataset.

### 3.3 Assigning Pharmacophore descriptors

A total of 9 pharmacophore descriptors were chosen on the basis of unique chemical features that play distinct role in binding to the DNA bases. The pharmacophores chosen were: aromatic, hydrophobic, hydrophilic, positively charged, negatively charged, and proline, cysteine, tryptophan and glycine were taken separately as they show unique characteristics.



| | | | |
|---|---|---|---|
| AROMATIC<br>{ F, Y } | R | PROLINE<br>{ P } | P |
| HYDROPHOBIC<br>{ A, V, L, I, M } | H | CYSTEINE<br>{C} | C |
| HYDROPHILIC<br>{ N, Q, S, T } | F | TRYPTOPHAN<br>{W} | W |
| POSITIVELY CHARGED<br>{ H, R, K } | O | GLYCINE<br>{ G } | G |
| NEGATIVELY CHARGED<br>{ D, E } | X | | |

**Fig.3.2** Notations for the pharmacophores assigned

Phenylalanine and tyrosine were grouped as aromatic; alanine, valine, leucine, isoleucine, and methionine were grouped as hydrophobic; asparagine, glutamine, serine, and threonine were grouped together as hydrophilic as these four amino acids have both hydrogen bond donor and acceptor groups in their side chain; histidine, arginine, and lysine were grouped as positively charged; aspartic acid and glutamic acid were grouped as negatively charged.

### 3.4 Building the pharmacophore matrix

To build the pharmacophore scoring matrix, we obtained the substitution frequencies for all the pairs of amino acids and also, the individual amino acid frequencies, which were used by Henikoff & Henikoff (Henikoff S. et al. 1992) to create the BLOSUM62 scoring matrix, from the BLOCKS database.

The amino acids were grouped into their respective pharmacophore and then the substitution frequencies were recalculated for each pharmacophore using an in-house developed python script. The individual frequencies of the amino acids were also recalculated for each pharmacophore.

Then, using the observed frequencies and the expected frequencies for each pharmacophore obtained, the log odds score for each pair of pharmacophores was calculated using an in-house developed python script.

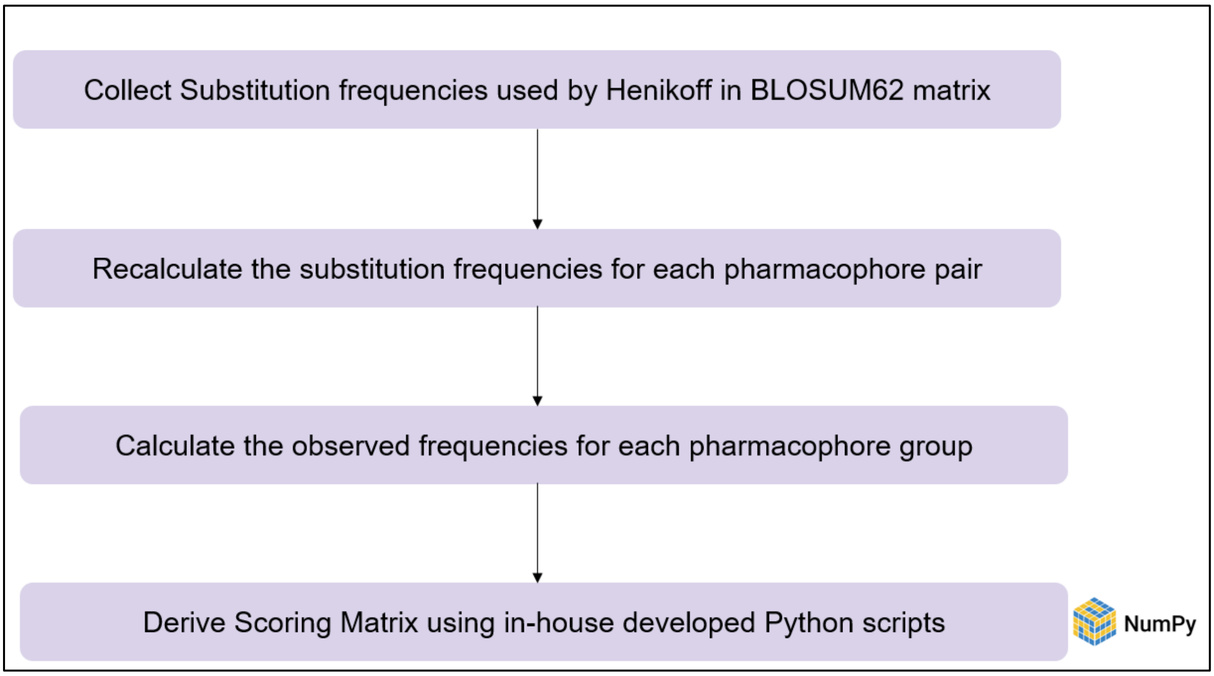

Collect Substitution frequencies used by Henikoff in BLOSUM62 matrix
Recalculate the substitution frequencies for each pharmacophore pair
Calculate the observed frequencies for each pharmacophore group
Derive Scoring Matrix using in-house developed Python scripts
NumPy

**Fig.3.3** Workflow for deriving the pharmacophore scoring matrix

### 3.5 Fetching the DNA interacting sequence of the protein

First, filtering was performed on the 263 homo-dimer proteins that were collected from NAKB database. In the filtering process, the duplicate proteins were removed from the set of sequences using a python script. There were several PDB IDs which had identical sequence as other IDs. These duplicates were removed by comparing the sequences and then getting a final FASTA file with only the unique sequences present in it.

Then the protein sequences whose corresponding DNA length is less than 8 nucleotides were removed. This was done because while we try to identify the DNA motifs, the MEME tool accepts DNA sequences only with length greater than or equal to 8 nucleotides. So, the small sequences were removed from the FASTA file using a python script.

After filtering, for each of the proteins, the corresponding PDB files were downloaded using a python script generated using ChatGPT. For each of the proteins, region which interacts with the DNA was identified using DNAProDB database by observing the interaction diagrams.

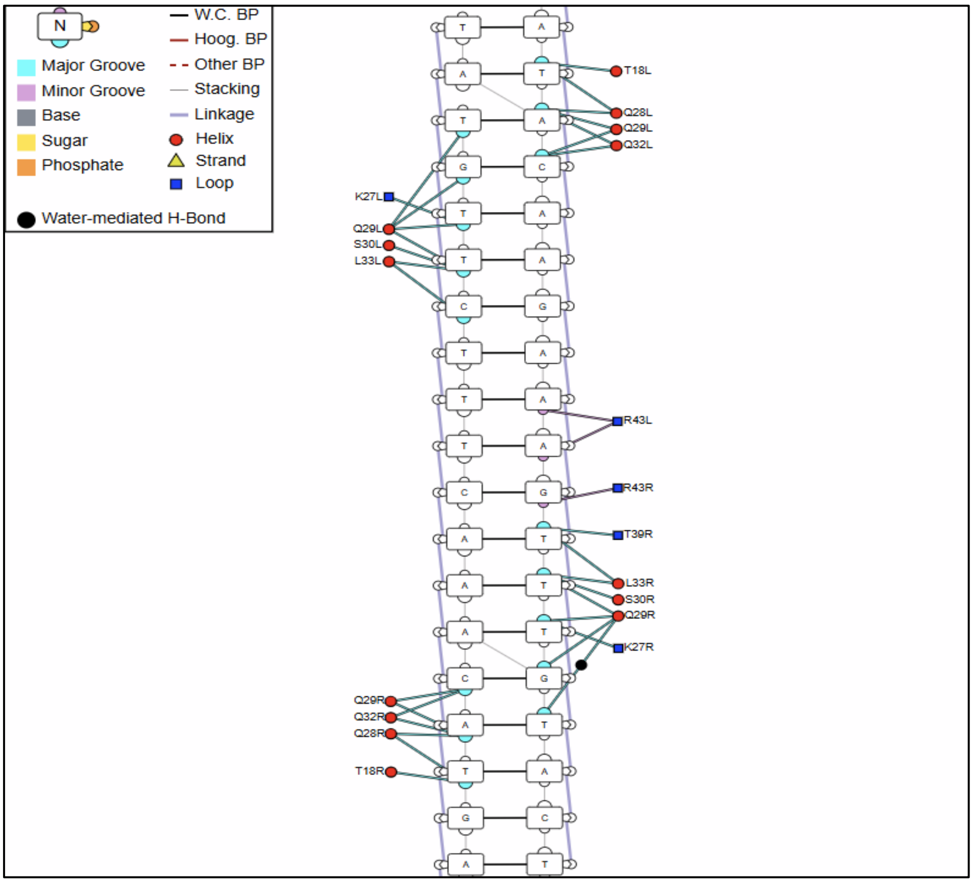
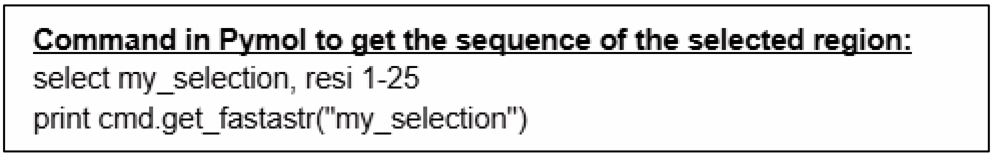


**Fig.3.4** Interaction diagram in DNAProDB showing the interactions between the amino acid residues and the DNA bases.

The sequence of the protein region which interacts with the DNA (the HTH domain) was extracted by visualizing the PDB file using the PyMol software. The interacting region was manually selected using mouse clicks and then using the command mentioned in figure 5, the sequence was obtained for the selected region. The sequences were then stored in the form of a FASTA file.



**Command in Pymol to get the sequence of the selected region:**

```
select my_selection, resi 1-25
print cmd.get_fastastr("my_selection")
```

**Fig.3.5** Command used for sequence extraction in pymol

### 3.6 Clustering of the protein sequences

The protein sequences were clustered based on their percentage sequence similarity using the CD-HIT tool, integrated within an in-house developed Python script for automation and batch processing. Various similarity thresholds were tested to identify the most informative clustering pattern. Among these, a 40% sequence identity threshold resulted in the most meaningful distribution, effectively grouping related sequences while maintaining sufficient diversity across clusters.

Upon execution of the script with the optimized threshold, CD-HIT generated multiple output FASTA files, each representing a distinct cluster and containing the corresponding set of protein sequences. This clustering approach facilitated downstream analyses by organizing the dataset into functionally or evolutionarily related groups.

### 3.7 Getting the consensus for each cluster

The consensus sequence for the list of sequences present in each cluster was obtained using an in-house developed python script. First, the sequences in each cluster were converted to the pharmacophore sequences using an in-house developed python script. Then, the sequences were aligned using offline ClustalW using the pharmacophore-based scoring matrix developed earlier. Four different alignment files were obtained, one for each cluster, in clustal format (.aln). Then the python script was used, where for each column in the alignment, it extracts the residues at that position across all sequences and counts the frequency of each residue. The most common residue in the column is identified as the consensus residue. A score is assigned based on the count of this consensus residue, with a score of 0 if the consensus residue is a gap. This residue and its corresponding score are appended to their respective lists. After processing all columns, the function calculates the cumulative consensus score by summing the scores of all positions, and finally returns the consensus sequence, the list of consensus scores, and the total cumulative score. We get 4 different text files containing results for each cluster.

### 3.8 Extracting DNA sequences

To retrieve the DNA sequences corresponding to the protein clusters, a Python script developed using ChatGPT was used. This script processes the FASTA files generated from the CD-HIT clustering step, parsing the PDB IDs embedded within the sequence headers. For each cluster, the script extracts and compiles these PDB IDs into separate CSV files. As a result, a total of four CSV files were generated, each representing the PDB IDs associated with a specific protein cluster.

Subsequently, another Python script utilizing the PDB RESTful API was employed to fetch the DNA sequences associated with each of the listed PDB entries. For every PDB ID, the script accessed the corresponding PDB webpage and extracted the DNA sequences from the reported protein-DNA complex structure. The output from this process consisted of four separate FASTA files, with each file containing the complete set of DNA sequences corresponding to the protein sequences in a particular cluster. This automated approach ensured accuracy, reproducibility, and efficient handling of large datasets.

### 3.9 Obtaining DNA motif for each cluster

To identify the DNA-binding motifs recognized by the helix-turn-helix (HTH) proteins in each cluster, the extracted DNA sequences were analysed using the MEME Suite—a widely used tool for discovering statistically significant sequence motifs.

Each of the four DNA sequence FASTA files, representing the four clusters, was individually uploaded to the MEME Suite web interface. The tool was configured to identify up to three motifs per cluster, optimized for typical transcription factor binding site lengths. The MEME analysis successfully identified a total of 12 DNA motifs across the four clusters. These motifs likely represent the conserved DNA elements targeted by the HTH domains within each sequence group, and they serve as a basis for further investigation into DNA-binding specificity, consensus site identification, and regulatory function.

# Chapter 4
# Results and Discussion

### 4.1 Sequences and Structures used in the study

The NAKB database contained a total of 14,320 annotated protein structures. Of these, 6,193 were identified as regulatory proteins. Among the regulatory proteins, 2,693 were classified as transcription factors, among which 1,579 were annotated specifically as DNA-binding transcription factors.

A subset of 765 DNA-binding transcription factors was found to contain helix-turn-helix (HTH) motifs. These were further categorized into five structurally distinct subtypes: Homeodomain (213 proteins), Lambda repressor-like (103 proteins), TetR family (38 proteins), Trp repressor (4 proteins), Winged helix/fork-head (402 proteins).

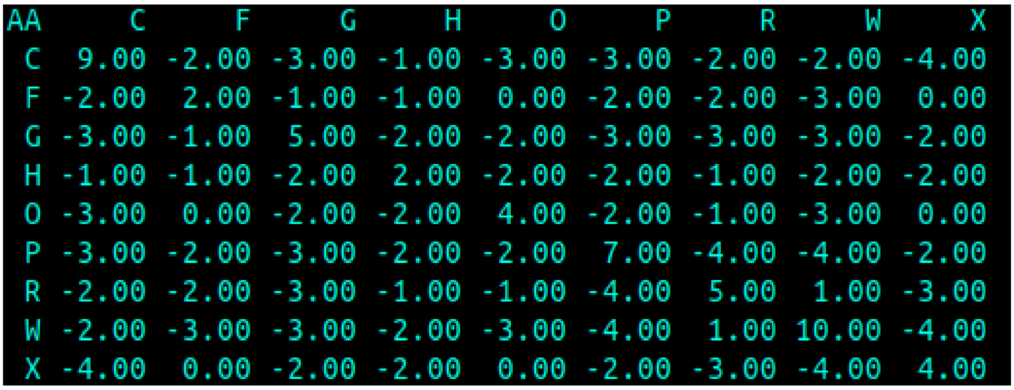
Out of the 765 HTH proteins, a total of 263 homo-dimeric HTH proteins were obtained using the API script. We chose to work specifically with the homodimer HTH proteins due to their unique DNA-binding properties. After this filtering process, a total of 155 HTH homodimer protein sequences were obtained which were used for the final analysis.

A total of 195 DNA sequences were collected across the four protein clusters: 71 sequences for Cluster 0, 38 for Cluster 1, 44 for Cluster 2, and 42 for Cluster 3. The number of DNA sequences exceeds the number of protein sequences because, in many cases, the PDB file contains sequences for both strands of the DNA. Since the homo-dimeric proteins often interact with both strands, each strand is represented separately in the dataset, resulting in a higher total count of DNA sequences than protein sequences.

### 4.2 Pharmacophore-based Scoring Matrix

A new pharmacophore scoring matrix was developed using in-house developed python script. The matrix contains the log-odds score for each pharmacophore pair mentioned in chapter 3.

Since Cysteine is unique, its high score suggests that it has distinct properties, likely due to its ability to form disulfide bonds. Negative scores with most other residues (e.g., C-X = -4.00, C-W = -2.00) indicate that Cysteine's interactions are highly selective.



```
AA     C     F     G     H     O     P     R     W     X
 C  9.00 -2.00 -3.00 -1.00 -3.00 -3.00 -2.00 -2.00 -4.00
 F -2.00  2.00 -1.00 -1.00  0.00 -2.00 -2.00 -3.00  0.00
 G -3.00 -1.00  5.00 -2.00 -2.00 -3.00 -3.00 -3.00 -2.00
 H -1.00 -1.00 -2.00  2.00 -2.00 -2.00 -1.00 -2.00 -2.00
 O -3.00  0.00 -2.00 -2.00  4.00 -2.00 -1.00 -3.00  0.00
 P -3.00 -2.00 -3.00 -2.00 -2.00  7.00 -4.00 -4.00 -2.00
 R -2.00 -2.00 -3.00 -1.00 -1.00 -4.00  5.00  1.00 -3.00
 W -2.00 -3.00 -3.00 -2.00 -3.00 -4.00  1.00 10.00 -4.00
 X -4.00  0.00 -2.00 -2.00  0.00 -2.00 -3.00 -4.00  4.00
```

**Fig.4.1** The pharmacophore scoring matrix derived using python script

Tryptophan (W) has the highest self-score (10.00). This suggests strong self-affinity, likely due to its bulky aromatic nature and involvement in π-π interactions. Negative interactions with negatively charged residues (X = -4.00) suggest that Tryptophan is not favourable in negatively charged environments.

Glycine is the smallest amino acid, providing structural flexibility. The matrix shows it has a high self-score (5.00) but does not interact well with other groups, likely because it lacks a side chain and contributes little to binding specificity. Thus, it has negative scores with all other groups.

Proline is a structural disruptor due to its rigid cyclic structure, which often introduces kinks in proteins. It has a high self-score of 7.00 but strong negative scores with positively charged (O = -4.00) and aromatic (R = -3.00) residues, which suggests that Proline disrupts stable charged/aromatic interactions.

Aromatic residues (R) have mixed interactions. The self-score is moderate, implying that π-π stacking is possible but not as strong as for Tryptophan (W). R-X score (-3.00) suggests that aromatic residues do not interact well with negatively charged residues.

Most interactions involving hydrophobic residues (H) are negative, meaning that hydrophobic residues do not significantly contribute to specificity. This aligns with their role in protein folding rather than direct molecular recognition.

There are several problems in this matrix that need to be fixed. There are some zero values such as for X-O, X-F, and O-F. This happened due to grouping of the amino acids which affected their observed frequency values used to calculate the log odds score.

## 4.3 Clustering of the protein sequences

Clustering of the final set of 155 homo-dimeric HTH domain sequences at a 40% sequence identity threshold resulted in the formation of four distinct clusters. The distribution of sequences among the clusters was as follows: Cluster 1 contained 55 sequences, Cluster 2 had 32 sequences, Cluster 3 comprised 35 sequences, and Cluster 4 included 33 sequences.

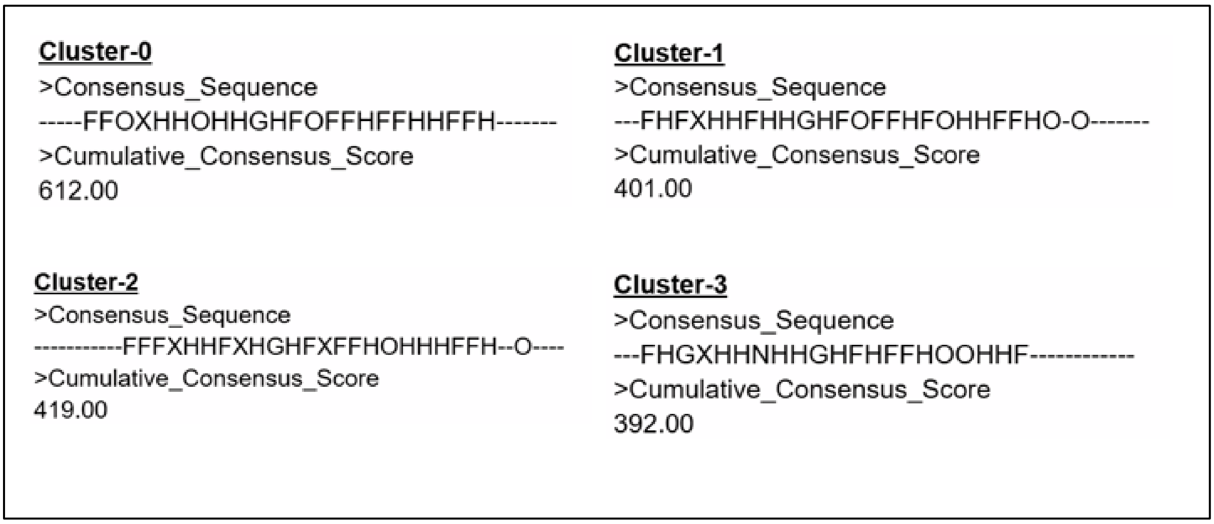
**Table 4.1** Number of protein sequences in each cluster.

| **Cluster number** | **Number of protein sequences** | **Average length** |
|---|---|---|
| 0 | 55 | 26 |
| 1 | 32 | 26 |
| 2 | 35 | 27 |
| 3 | 33 | 23 |

## 4.4 Obtaining the consensus sequence

The pharmacophore consensus sequence for each cluster was obtained. Figure 6 shows the pharmacophore consensus sequence and the consensus score for each cluster obtained.

After obtaining the 4 consensus sequences, they were again aligned with each other in ClustalW using the pharmacophore scoring matrix developed earlier. A final consensus sequence was obtained which represents the HTH homodimers.



```
Cluster-0
>Consensus_Sequence
-----FFOXHHOHHGHFOFFHFFHHFFH-------
>Cumulative_Consensus_Score
612.00

Cluster-1
>Consensus_Sequence
---FHFXHHFHHGHFOFFHFOHHFFHO-O-------
>Cumulative_Consensus_Score
401.00

Cluster-2
>Consensus_Sequence
-----------FFFXHHFXHGHFXFFHOHHHFFH--O----
>Cumulative_Consensus_Score
419.00

Cluster-3
>Consensus_Sequence
---FHGXHHNHHGHFHFFHOOHHF------------
>Cumulative_Consensus_Score
392.00
```

**Fig.4.2** The consensus sequences and the consensus scores for the clusters.

The 3 regions of the HTH domain- the stabilizing helix, the turn, and the recognition helix were marked on the consensus sequence by visualizing the sequences on PyMol. Figure 7 shows the final consensus sequence with the three regions of the HTH domain marked distinctively.

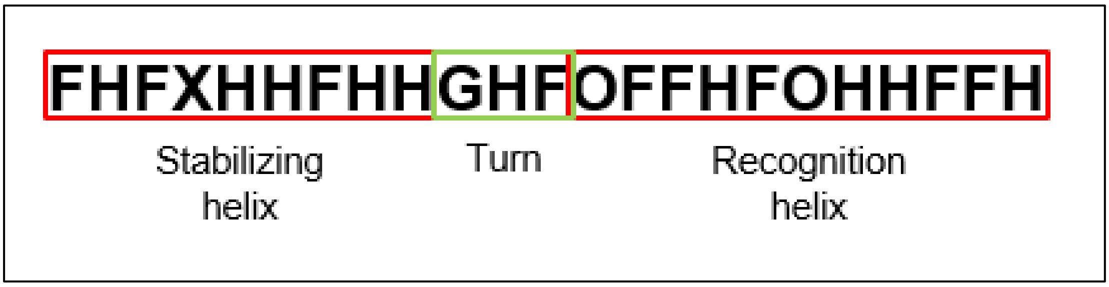


**Fig.4.3** Final consensus sequence with the three structural regions highlighted.

### 4.5 Obtaining the DNA motifs for each cluster

The DNA motifs for each cluster were obtained using MEME suite and 3 motifs were obtained for each cluster. On an average, each of these motifs were derived using 8 DNA sequences for each. Out of the 12 DNA motifs, 4 best motifs were chosen (one motif for each cluster). The best motif for each cluster was chosen by checking the number of DNA sequences involved and the pattern that was observed in the motifs. Figure 8 shows the final 4 motifs which were obtained.

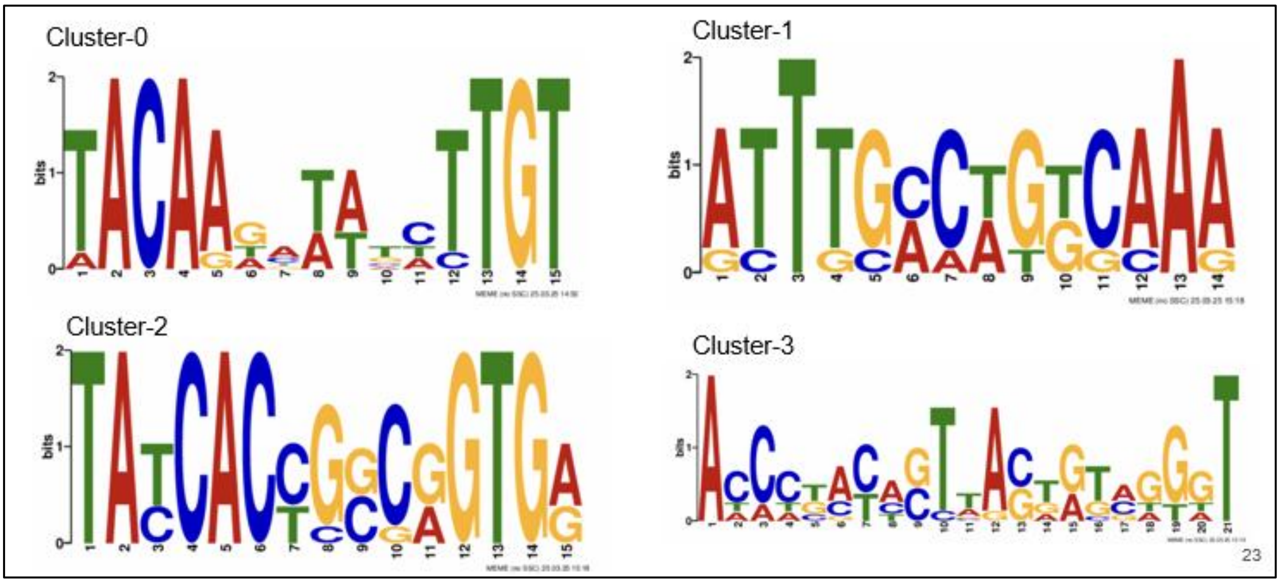


**Fig.4.4** DNA motif for each cluster with the cluster name.

**Table 4.2** Number of DNA sequences involved in generation of the motifs for each cluster.

| Cluster number | No. of DNA sequences involved in motif |
|---|---|
| 0 | 8 |
| 1 | 6 |
| 2 | 9 |
| 3 | 11 |

## 4.6 Discussion

The analysis of protein–DNA interactions, particularly in the context of homo-dimeric helix-turn-helix (HTH) transcription factors, provides critical insights into the regulatory mechanisms governing gene expression. Our study demonstrates how systematic data mining from curated databases like NAKB, along with computational tools such as CD-HIT, MEME Suite, and custom-built Python scripts, enables comprehensive exploration of structural and sequence-level diversity among HTH proteins. The identification of 155 high-confidence homo-dimeric HTH proteins and their subsequent classification into four distinct sequence-based clusters revealed notable variation in DNA-binding preferences, as evidenced by the distinct motifs identified in each cluster. These results underscore the functional heterogeneity even among proteins sharing the same structural fold, which reflects evolutionary diversification in target DNA recognition.

The pharmacophore scoring matrix plays a critical role in solving the motif detection problem by enabling biologically meaningful comparisons of protein sequences transformed into pharmacophore-based representations. Unlike traditional matrices like BLOSUM or PAM that rely on evolutionary substitution rates, this matrix captures the chemical compatibility of pharmacophore pairs using log-odds scores derived from their observed co-occurrences. This allows for alignment of pharmacophore sequences where chemically similar but non-identical residues are treated as functionally equivalent, which is especially valuable in diverse but functionally conserved protein families like HTH homodimers. By guiding the alignment process with scores that reflect actual biochemical properties, the matrix helps identify consensus regions that are not just sequence-conserved but chemically conserved, revealing potential functional motifs. These consensus sequences, when evaluated using the matrix, yield high-confidence positions that likely contribute to DNA binding.

Homo-dimeric HTH proteins are known for their symmetric DNA-binding behaviour, where two identical subunits interact with palindromic or symmetric DNA sequences, leading to enhanced binding specificity and stability. This symmetry simplifies the analysis of protein-DNA interactions, making it easier to identify conserved motifs and structural features. Moreover, homodimers tend to have more consistent pharmacophore and sequence patterns across their DNA-binding domains, which aids in the generation of high-quality consensus sequences and facilitates accurate clustering.

When we considered only the pharmacophore sequence of the recognition helix of the HTH domain, it was observed that there were several conserved sites in the patterns. The patterns differed only at 3 sites (first, fifth, and sixth) of the recognition helix. The comparison of the recognition helix patterns has been shown in table 2.

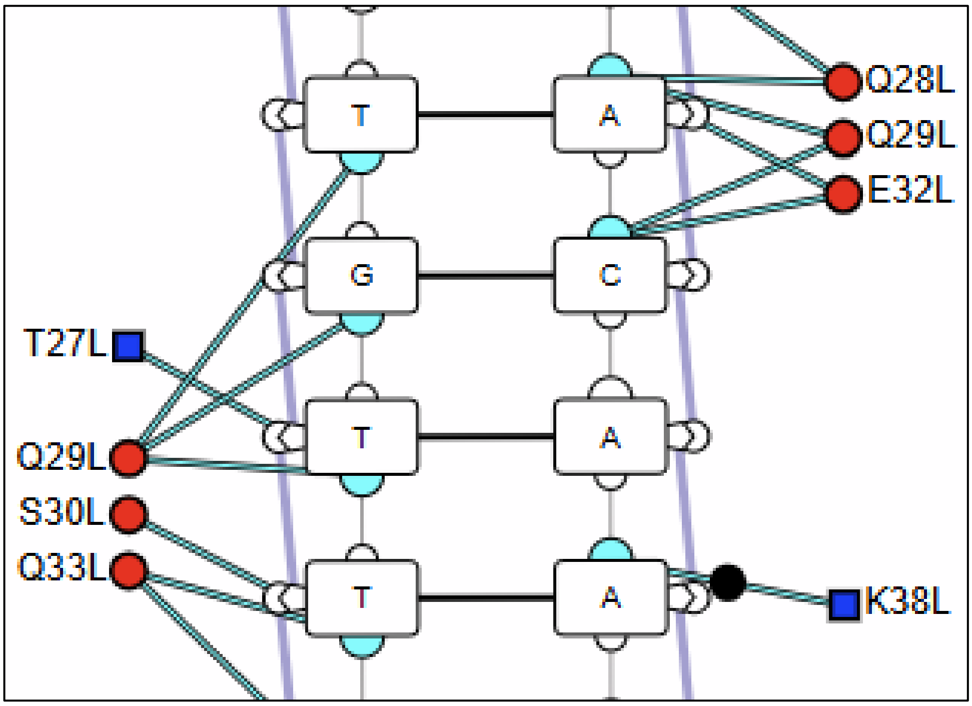
**Table 4.3** Variable region in the recognition helix patterns among the 4 clusters. The variable sites have been highlighted in red.

| Cluster number | Recognition Helix pattern |
|---|---|
| 0 | **O**FFH**FF**HHFFH |
| 1 | **O**FFH**FO**HHFFH |
| 2 | **X**FFH**OH**HHFFH |
| 3 | **H**FFH**OO**HHFFH |

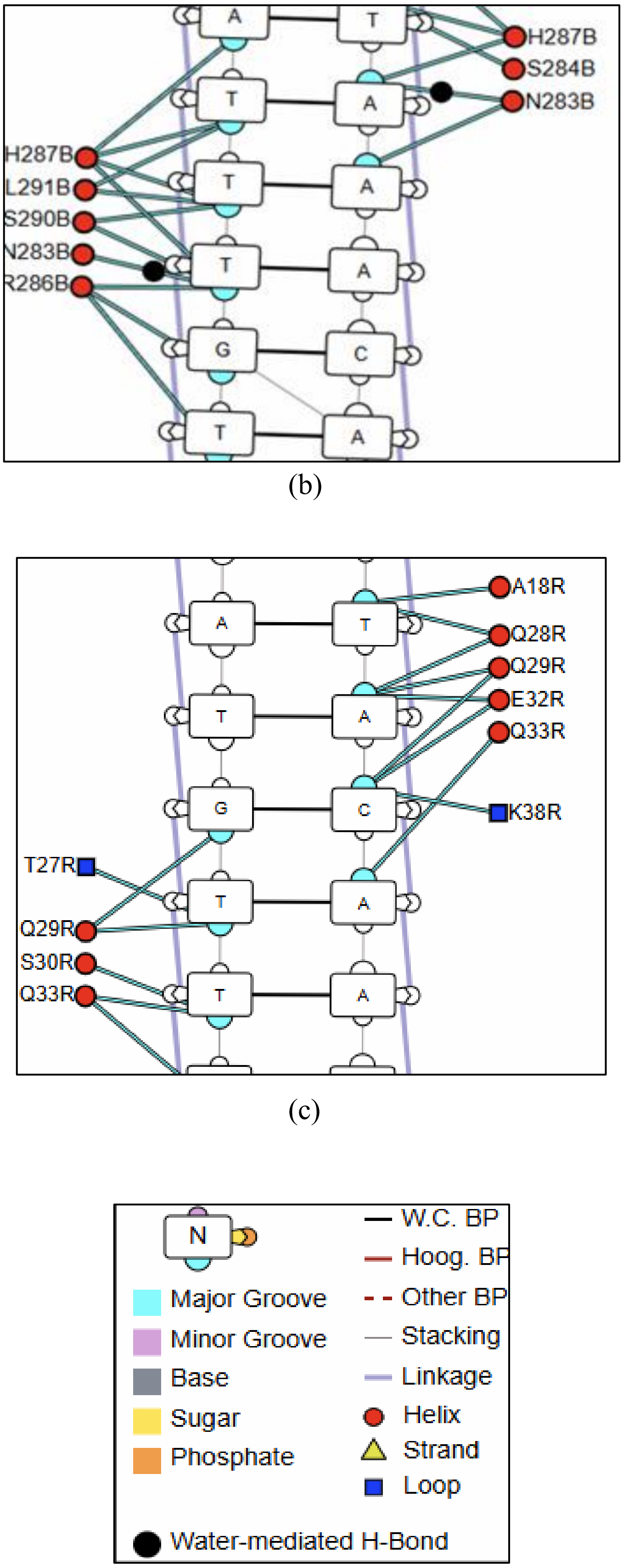
It was observed that the HTH domain has an average length of 23, where the stabilizing helix is 9 residues, the turn is 3 residues, and the recognition helix is 11 residues long respectively. The stabilizing helix mostly consists of hydrophobic residues (5 out of 9). The turn has at least one glycine present which provides the required flexibility. The recognition helix mostly consists of hydrophilic and charged residues.

The DNA motifs obtained show broken palindromic patterns, which shows that both the HTH domains of the homodimer proteins show identical binding to the DNA. Palindromic patterns in the DNA-binding sites of HTH homodimer proteins are highly significant because they reflect the structural symmetry of the protein itself. These palindromic DNA sequences allow each subunit of the homodimer to interact with one half of the binding site in a mirror-image manner, facilitating strong and specific binding. This structural complementarity enhances binding affinity, increases the stability of the protein-DNA complex, and contributes to precise regulation of gene expression.

On analysing the interaction diagrams of the DNA-protein complexes using DNAProDB database, there were several patterns observed in the positions of the amino acid residues that interact with the DNA bases. It was observed that in majority of the cases, there can be seen a conserved pattern in the spacing between the interacting residues. As shown in figure 11(a), the 29th, 30th and 33rd residues interact with the DNA bases on one DNA chain. In figure 11(b), 287th, 290th, and 291st residues, or 283rd, 286th, and 287th residues interact with the bases on one DNA chain. Similarly, in figure 11(c), 29th, 30th, and 33rd residues interact with the DNA bases on one chain. All these interactions show a common pattern which is same gapping between the residues at the mentioned positions. Among the three residues, two are adjacent residues and two are separated by 2 residues. Apart from this pattern, in many cases a residue at a position 10 residues away from the previously mentioned pattern was also observed, such as in the case of figure 11(c) where the 18th, 28th, 29th, and the 32nd residues interact with the same chain of DNA.



(a)



(b)

(c)

(d)

**Fig.4.5** Interaction diagrams for: (a) 1RPE (b) 2OR1 (c) 3G73 from DNAProDB. (d) index for the diagrams

# Chapter 5
# Summary and Conclusion

## 5.1 Summary

This project focused on helix-turn-helix (HTH) proteins, a key group of DNA-binding proteins that help control gene expression in many organisms. The main goal was to understand how these proteins recognize and bind to specific DNA sequences. A combination of bioinformatics tools and custom scripts was used to analyse protein-DNA interactions, identify conserved motifs, and develop a new pharmacophore-based scoring matrix.

We started by collecting data from the Nucleic Acid Knowledge Base (NAKB), which contains detailed information about proteins that interact with DNA. From 14,320 protein structures, we identified 765 that had the typical HTH DNA-binding domains. We then narrowed this down to 155 proteins that form homodimers (pairs of the same protein) and bind symmetrically to palindromic DNA sequences. Homodimers were chosen because their symmetric DNA-binding behaviour simplifies the study of recognition patterns.

To better understand the chemical properties influencing DNA binding, a pharmacophore-based scoring matrix was developed. This matrix grouped amino acids based on their chemical features and assigned scores to their interactions. This approach helped in identifying functionally important regions in the HTH domain, particularly the recognition helix, which directly contacts DNA.

The protein sequences were clustered into four groups based on sequence similarity, and a consensus sequence was derived for each cluster. The HTH domain was found to have an average length of 23 residues, with distinct regions: a stabilizing helix (9 residues), a short turn (3 residues), and a recognition helix (11 residues). The stabilizing helix was mostly hydrophobic, while the recognition helix contained more charged and hydrophilic residues, crucial for DNA interaction.

DNA sequences bound by these proteins were analysed using the MEME Suite, revealing palindromic or near-palindromic motifs, consistent with the symmetric binding of homodimers. Additionally, a recurring pattern in residue spacing was observed, where specific positions (e.g., 29th, 30th, and 33rd residues) frequently interacted with DNA bases, suggesting a conserved binding mechanism.

### 5.2 Conclusion

In this project, we successfully analysed the structure and function of homo-dimeric HTH proteins, uncovering key patterns in their DNA-binding mechanisms. By clustering sequences and deriving consensus patterns, the study revealed that the recognition helix has conserved chemical properties critical for DNA interaction. The development of a pharmacophore-based scoring matrix provided a new way to evaluate these interactions, highlighting the importance of specific residues in binding. Additionally, the discovery of palindromic DNA motifs confirmed the symmetric binding nature of homo-dimeric HTH proteins.

One of the major findings was the consistent spacing between DNA-binding residues, suggesting a common structural mechanism across different HTH proteins. These insights enhance our understanding of gene regulation and open new possibilities for designing molecules that can modulate protein-DNA interactions. Overall, this study advances our knowledge of HTH proteins and provides a foundation for further research in genetic and therapeutic applications.

# Chapter 6
# Future Work

One key area for future research involves developing computational methods to accurately identify and quantify broken palindromes in DNA sequences recognized by HTH proteins. While perfect palindromes are well-studied, many HTH-binding sites exhibit slight asymmetries or mismatches in their inverted repeats. A systematic approach to detect and score these imperfect palindromes would provide deeper insights into how HTH dimers tolerate sequence variations while maintaining binding specificity. This could involve creating new algorithms that consider both sequence symmetry and the structural flexibility of HTH domains, potentially leading to improved prediction of non-canonical binding sites in genomic data.

Additionally, the sequences separating the two halves of palindromic binding sites (spacer regions) remain understudied, despite their potential role in modulating DNA-protein interactions. Future work should analyse these spacer regions to determine whether their length, sequence composition, or structural properties influence binding affinity or specificity. High-throughput binding assays combined with machine learning could reveal whether certain spacer patterns correlate with stronger or weaker HTH-DNA interactions, which would refine our understanding of how these proteins recognize their targets.

Finally, there is a need for more robust computational tools to predict DNA sequences that bind to HTH proteins directly from primary sequence data. Current methods often rely on known motifs or structural data, limiting their applicability to uncharacterized HTH proteins. Future efforts could integrate the pharmacophore-based scoring matrix developed in this study with deep learning models trained on experimentally validated binding sites. Such a tool would be valuable for annotating putative regulatory elements in newly sequenced genomes and could aid in the design of synthetic HTH proteins with customized DNA-binding specificities for biotechnological applications.

# References

1) Brennan, R. G. (1992). DNA recognition by the helix-turn-helix motif: Current Opinion in Structural Biology 1992, 2: 100…-108. *Current Opinion in Structural Biology*, *2*(1), 100-108.

2) Pellegrini-Calace, M., & Thornton, J. M. (2005). Detecting DNA-binding helix–turn–helix structural motifs using sequence and structure information. *Nucleic acids research*, *33*(7), 2129-2140.

3) Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F., & Matthews, B. W. (1982). Structural similarity in the DNA-binding domains of catabolite gene activator and cro repressor proteins. *Proceedings of the National Academy of Sciences*, *79*(10), 3097-3100.

4) Anderson, W. F., Ohlendorf, D. H., Takeda, Y., & Matthews, B. W. (1981). Structure of the cro repressor from bacteriophage λ and its interaction with DNA. *Nature*, *290*(5809), 754-758.

5) McLaughlin, W. A., & Berman, H. M. (2003). Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif. *Journal of molecular biology*, *330*(1), 43-55.

6) Rosinski, J. A., & Atchley, W. R. (1999). Molecular evolution of helix–turn–helix proteins. *Journal of molecular evolution*, *49*, 301-309.

7) Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature*, *353*(6346).

8) Luscombe, N. M., Austin, S. E., Berman, H. M., & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome biology*, *1*, 1-37.

9) Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919

10) Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., & Lu, P. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science (New York, N.Y.)*, *271*(5253), 1247–1254.

11) Busby, S., & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). Journal of molecular biology, 293(2), 199–213.

12) Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y., & Matthews, B. W. (1982). The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. *Nature*, *298*(5876), 718–723.