**SCOPE**

The scope of this project involves the identification and analysis of anchor segments within transcription factor binding sites (TFBS) using data from both **in vitro** and **in vivo** experiments. Specifically, we will utilize datasets from UniProbe and CIS-BP databases, which include **in vitro** data from protein-binding microarrays (PBM) and systematic evolution of ligands by exponential enrichment (SELEX), as well as **in vivo** data from chromatin immunoprecipitation sequencing (ChIP-Seq). Our focus will be on different types of transcription factors (TFs)—including monomeric and dimeric TFs—that bind to both gapped and ungapped sequences.

Building upon the novel alignment algorithm developed last semester, we aim to enhance it to better account for the best binding sites. This involves moving away from the simplistic alphabetical alignment of substrings and accommodating dimer binding in both continuous and gapped forms. By refining the algorithm in this manner, we will progress toward our primary objective: designing drugs that restrict TF binding by targeting essential anchor segments required for gene expression. Furthermore, a web server will be developed to allow users to upload their own transcription factor data, run the anchor residue finder and visualizer, and interact with a 3D model viewer that visualizes the binding of the transcription factor on DNA with the anchor residues highlighted.

**Constraints:**

The project will be constrained by the availability and quality of data from UniProbe and CIS-BP. Resource limitations, including computational resources and time constraints, may impact the scale and complexity of our analysis.

**Deliverables:**

* A comprehensive report detailing the methodology, findings, and conclusions of the project.
* An enhanced algorithm applicable to any chosen dataset.
* A publicly accessible web server based on Django (Python) to deploy our tool.

**PLAN OF WORK**

Through this project, we will focus on the following key aspects:

1. **Data Procurement and Preparation**

We will procure relevant data from UniProbe, CIS-BP, and specific research papers for a variety of transcription factors, including monomeric and dimeric TFs such as CEBPa, CEBPb, FOS, and other glucocorticoids. This data will encompass both **in vitro** experiments (PBM, SELEX) and **in vivo** experiments (ChIP-Seq).

1. **Understanding Experimental Data Acquisition**

A thorough understanding of how data is acquired through experiments like ChIP-Seq, SELEX, and how these differ from the PBM experiments of UniProbe is crucial. This knowledge will allow us to adapt our algorithm to accurately represent and analyze the different types of data.

1. **Enhancing the Alignment Algorithm**

The current algorithm sorts substrings of the same length alphabetically, which is not biologically meaningful. We will enhance the algorithm by:

* Counting the number of substring matches after alignment.
* Selecting the alignment that results in the maximum number of common substrings with the top seed motif.
* Accommodating dimer binding in both continuous and gapped forms.

This enhancement may increase computational time, so we will explore optimizing the algorithm through parallel processing and logical improvements.

1. **Working on Web Server**

We will be creating a Django based web server which will support background processes. The idea is that the user will enter the TF for which the anchor residues will be generated, and the server will give them a job code. Upon entering the job code, the results of the anchor residue algorithm can be retrieved.

1. **Research Paper Writing**

In this paper, we will present our novel research efforts aimed at developing a powerful tool to aid the design of drugs that specifically target transcription factor (TF) binding sites.

By integrating these key steps and methodologies, our approach seeks to advance the understanding and regulation of gene expression. This tool has the potential to significantly contribute to future breakthroughs in targeted therapies, offering new strategies for manipulating gene activity in disease contexts.