

SCOPE

The scope of this project encompasses the identification and analysis of anchor segments within transcription factor binding sites (TFBS) using UniProbe, CIS BP or TF specific paper data and the novel alignment algorithm developed in the last semester. The algorithm will be further enhanced - first to account for best binding site by doing away with alignment of substrings merely alphabetically and to accommodate dimer binding - continuous and with gaps. By doing this, it will help us make progress towards the primary objective, which is to be able to design drugs that restrict binding by targeting anchor segments which are required for gene expression.

Constraints:

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

Deliverables:

I'll be able to produce a report detailing the methodology, findings, and conclusions of the project. My code maybe implemented to any chosen dataset. Additionally, I'll be working on a web server based in Python, which can be made accessible publicly.

PLAN OF WORK

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

(1) Data Procurement and Preparation:

I'll be procuring relevant data from the UniProbe, CIS BP or paper-specific database for a variety of dimer binding TFs (eg: CEBPa, CEBPb, FOS, and other glyocorticoids).

(2) Understanding how the Data is Acquired:

I'll be understanding these experiments: CHIP Seek, SELEX and how it is different from the PBM experiments of UniProbe, and see if my algotihm can be adapted to represent the changes.

(3) Enhancing Alignment Algorithm:

Until now, the algorithm sorts the substrings of same length by alphabets. However, this doesn't make sense from a biological perspective. I'll be working on counting the number of substring matches after alignment, and based on after which alignment, there are maximum number of common substrings with the top seed motif, that alignment will be chosen. This might result in higher running times, so optimising the algorithm through parallelism and changes in logic, will have to be considered.