Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2024-2025

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	http://sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

Prediction of DNA binding motifs based on protein structure

Keywords:

DNA binding motifs; Gene-regulatory network; transcription factors; Homology Modelling

Summary:

The process begins by segmenting the TF sequence into fragments, which can be structurally modeled using homolog templates. Models from different templates with similar sequence fragment are clustered, and the largest TF sequence aligning with each cluster is retained. Remaining fragments are stored for further analysis. A PWM is predicted for each fragment. Here's how PWM prediction for a fragment works: 1. Homolog Search: BLASTP is used to identify homologs of the fragment within JASPAR, CisBP and a database of sequences with known structures in complex with DNA. 2. PWM Prediction: PWM prediction is performed using a combination of the nearest-neighbor and ModCRE approaches. Predicted PWMs from both methods are clustered based on similarity using TOMTOM. For the same fragment we may have more than one cluster, corresponding to different PWMs. 3. Average PWM: We compute a weighted-average PWM for each cluster (CPWM). Weights for PWMs within a cluster are determined based on the similarity between the fragment's sequence and the homolog's sequence used as a template. Weights are obtained with an AI classifier (Risk 1). 4. AI Classifier: AI classifiers are developed to obtain the weights and to rank the reliability of CPWMs, potentially providing multiple predictions (we name as primary selection the best predicted candidate and all other as secondaries). 5. Ab Initio Modeling: If no close homologs are found by the nearest neighbor method or no templates are available for modeling the DNA complex, an ab initio structure modeling approach is employed (subtask 1.1). Subtask 1.1: ab initio model building of a TF-DNA structure. The objective involves the ab initio construction of a TF-DNA complex structure when there are no suitable homologous templates available. We will obtain the structure of the protein with AlphaFold2 65 or from the datasets of predicted structures59. To complete the model with Alphafold2, we will search the most similar structure(s) among all structures of proteins bound with DNA and superpose the proteins. The superposition will locate the structure of the DNA molecule and this will be sufficient to reconstruct the bounded complex (Risk 2) Risks: Risk 1: determining weights for averaging PWMs within a cluster based on sequence similarity between the fragment and the homolog's sequence used as a template might not result in an accurate PWM. It's possible that some PWMs perform better in certain positions of the matrix, while others excel in different positions. Alternative plan: To address this challenge, a more sophisticated approach can be employed, such as implementing position-specific weights for each position of the PWM. This could involve using methods like the similarity-regression approach to refine the PWM construction process. Risk 2: In subtask 1.1, we require the structural similarity between the modelled TF and the template bound with DNA to apply a superposition. However, if no similar structures to be used as template are found in the database, then we need to apply another approach. Alternative plan: We will use two possible protocols: 1) we will use RosettaFolDNA to construct the protein-DNA complex; 2) we'll proceed using a docking approach. For the docking approach: First, the most probable amino-acids to bind DNA will be predicted (i.e with DBSI). Then, a standard B-DNA conformation will be built with X3DNA. Finally, both molecules will be docked using any of the methods available: with NPDock, pyDockDNA, HDOCK or an in-house protocol using FTDock).

References:

Meseguer, A. et al. On the prediction of DNA-binding preferences of C2H2-ZF domains using structural models: application on

human CTCF. NAR Genom Bioinform 2, lqaa046 (2020) Fornes, O. et al. Automated structure-based learning to model co-operativity and protein-DNA interactions in cis-regulatory modules. bioRxiv 2022.04. 17.488557 (2022)

Expected skills:

Python programming skills

Possibility of funding:

No

Possible continuity with PhD:

To be discussed

Comments:

Hybrid work remote/presenial allowed