Modeling of ssDNA binding protein structures

Maanav Jariwala, OUR Summer Research Scholar
Dr. Jun-tao Guo, Department of Bioinformatics and Genomics

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Abstract

Single-stranded DNA (ssDNA) binding proteins are crucial for biological processes including replication, transcription, and genomic integrity. We can better comprehend the function of the proteins and their interactions with ssDNA with structures of ssDNA binding proteins. However, the number of known ssDNA binding protein structures is still very small, despite recent advances in experimental protein determination methods. Lately, the implementation of AlphaFold2, a state-of-the art, artificial intelligence-based method, has led to a breakthrough in protein structure prediction. In this project we used the predicted ssDNA binding protein structures from AlphaFold2, on a dataset of annotated non-redundant ssDNA binding proteins. We investigated the quality of the predicted models and the relationship between protein size and model quality score

Introduction

- DNA replication, repair, and recombination are among the biological processes in which single-stranded DNA binding proteins (SSBs) play a crucial role.
- Knowledge of ssDNA structures is crucial for understanding processes. Inappropriate recognition can result in chromosomal damage leading to processes like cancer, senescence, or cell death.
- More high quality of ssDNA binding protein structures can help improve our understanding of protein-ssDNA interactions

Objectives

- The objective of this study is to predict the annotated ssDNA binding protein structures and investigate their model quality

Methods

- Use a tab-delimited file containing primary accession IDs of SSBs clustered at a cutoff of 35% sequence similarity to acquire each PDB file of predicted structures by AlphaFold2. (Image 1)
- Run a script to take each accession id and search the ebi model database predicted by AlphaFold2(Figure 1)
- Run a script that finds the average quality score of each PDB file
- Run a script that categorizes each PDB file into a specific quality score level group. (Figure 2)

```python
#!/usr/bin/env python3
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC

PATH = "/Users/maanav/Desktop/chromedriver"

openfile = open("ssb_clustered_35.tsb")
file_contents = openfile.readlines()

driver = webdriver.Chrome(PATH)

accession_list = []
for x in range(1,len(file_contents)):
    accession_list.append(file_contents[x].strip())

structure_list = []
nonstructure_list = []

for ID in range(0,len(accession_list)):
    if driver.get("https://alphafold.ebi.ac.uk/files/AF-" + str(accession_list[ID]) + "-F1-model_v2.pdb"):
        structure_list.append(accession_list[ID])

    else:
        nonstructure_list.append(accession_list[ID])

print("Structured accession ID list:",structure_list)
print("UNstructured accession ID list:",nonstructure_list)
```

Image 1: Python script to get figure 1 results

Results

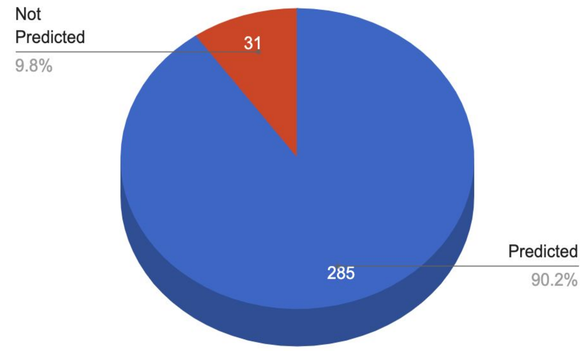


Fig. 1: Majority of the non-redundant ssDNA binding proteins (90.2%) in our dataset have predicted models by AlphaFold2 at alphafold.ebi.ac.uk

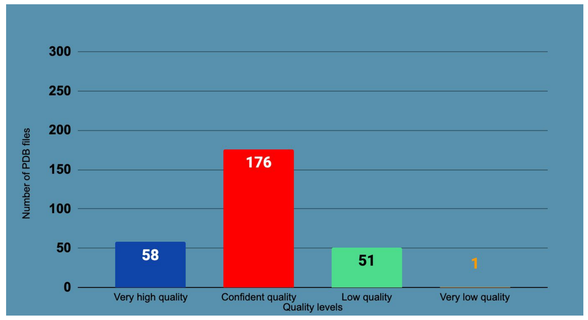
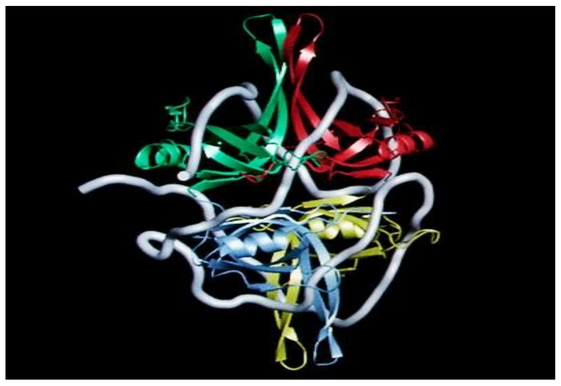


Fig. 2: of the models 58 (20.2%) are in the very high confident category, followed by the majority 176 (61.5%) being in the confident category, and 51 (17.5%) in the low quality, finally 1 model (0.8%) is in the very low category.

Fig. 3: Scatterplot shows the relationship between the quality score and protein size for all the 286 protein PDB files.

- The trend in the comparison of protein size to the quality score for each protein, is that there is a better quality score in smaller proteins. (Figure 3)



Phs Org (2009)

Discussions

- Majority of the ssDNA binding proteins have predicted structures by AlpfaFold2 at EB
- Smaller proteins tend to have better model quality
- Finally, high quality protein models will be compared with existing structures for novel fold identification

References

Staff, S. X. (2009, October 21). Single-stranded DNA-binding protein is dynamic, critical to DNA repair. Single-Stranded DNA-Binding Protein Is Dynamic, Critical to DNA Repair. https://phys.org/news/2009-10-single-stranded-dna-binding-protein-dynamic-critical.html

NCBI - WWW Error Blocked Diagnostic. (2013, July 2). Single-Stranded DNA-Binding Proteins: Multiple Domains for Multiple Functions. https://pubmed.ncbi.nlm.nih.gov/23823326/