# Exploration of C3UPD2 Protein in Escherichia coli

Daniel Forrester
C1028045

## Contents



## Introduction

One protein that piques my curiosity the most out of the list of five to choose from is C3UPD2 · C3UPD2_ECOLI, a resident of the well-known Escherichia coli (E. coli) bacterium. E. coli's significance in common diseases adds an intriguing dimension to this exploration. What captivates me the most is the apparent underrepresentation of C3UPD2, marked by its annotation level of 1/5 in the UniProt database. This low annotation level points to a scarcity of information surrounding this protein and may indicate that for the little information we know may even be incorrect or inaccurate.

The objective of this report is to explore the uses, C3UPD2. My approach is based on well-established application of bioinformatics tools to scrutinise the existing record in UniProt, ultimately aiming to gain insights regarding this accuracy of the documented protein. In this report, I will undertake a conclusive evaluation of C3UPD2, developing insights into its fundamental identity, functional attributes, homology to develop and document data of this scarcely recorded protein.

## Explanation of the methods and tools used for the analysis

Sequence alignment stands as a vital technique for comparing and dissecting the nuances for protein structures. And is mainly used to pinpointing shared regions of resemblance, scrutinising genetic mutations, tracing evolutionary lineage of homology, and possibly discovering or linking functional elements of such sequences.

To begin I will use the UniProt database as it'll serve as the main source for the protein's sequence information, it's Primary Structure, Secondary Structure, Function and Homology which would be pivotal for true understanding of what's already know about the protein. Then to create 3D models of C3UPD2, I will then use online servers, precomputed models from databases, and the software like

PyMol and Modeller. These models will be then visually compared in PyMol to assess their quality and choose the most suitable one.

The prediction of the protein's secondary structure will be carried out to gain insights into its conformation. Finding a protein's secondary structure reveals its local structural elements like alpha-helices, beta-sheets, or loops [R1], this information is crucial for understanding its function.

The analysis will use molecular biology databases like EMBL, GenBank, and DDBJ for nucleic acid information whilst BLAST and UniProt were used for protein sequences. The accuracy of the UniProt record for C3UPD2 was verified by cross referencing the data from other databases like NCBI, Swiss-Pro and Ensembl to ensure consistency and reliability. This approach would allow me to analyse the protein's structure, shedding light on its potential functions and significance in E. coli.

## Identifying and analysing homologous proteins

C3UPD2 · C3UPD2_ECOLI is found in the bacterium Escherichia coli (E.coli) which is a common and well-studied model organism and knowing the homology of C3UPD2's host organism, although not necessarily equivalent, can show clues to the homology of the protein itself.

But the relationship between organisms can be assessed based on their "taxonomic classification" and in this case, Aliivibrio fischeri (strain MJ11) (formerly known as Vibrio fischeri) is more closely related to Escherichia coli (strain K12). Escherichia coli and Aliivibrio fischeri both belong to the class "Gammaproteobacteria" and are within the same bacterial phylum, Proteobacteria. While they are not closely related in terms of their genus (Escherichia vs. Aliivibrio), they share a closer taxonomic relationship compared to the other organisms listed, which belong to the genus Vibrio (Vibrio cholerae) or are uncultured bacteria.

In addition to what was found previously about the fluorescence of C3UPD2, Aliivibrio fischeri (strain MJ11) exhibits bioluminescent properties. Emitting visible light as a result of a chemical reaction involving luciferase, luciferin, and oxygen. It is not known for fluorescence but rather bioluminescence, which is used for communication and symbiotic interactions in marine environments.

## BLAST Report

After BLASTing, C3UPD2 · C3UPD2_ECOLI is 100% identical with 31 other sequences, 12 of with originating from Mus Musculus (house mouse), and the rest in: Escherichia coli, synthetic constructs, Aequorea victoria, Burkholderia cenocepacia, Bacillus subtilis, Saccharomyces cerevisiae, Plasmodium falciparum, Avena sativa, and several mammalian species like Homo sapiens and Mus musculus as stated before.

But most importantly "Chain S, Green fluorescent protein [Escherichia coli]" is both 100% sequence identity, the same alignment length and has a low e-value of 5e-04, which would imply that these are both the same protein. This suggests that it's a variant of Green Fluorescent Protein (GFP) found in the bacterium Escherichia coli (E. coli) and GFP itself is a naturally occurring protein (independent from ecoli) that exhibits green fluorescence when exposed to ultraviolet or blue light. GFP and its variants are frequently used as molecular markers to study gene expression, protein localisation, and other cellular processes. They are also used in various bioimaging techniques and fluorescence-based assays

To visualise some of C3UPD2 · C3UPD2_ECOLI I created a phylogenetic tree [F1] using a selected set of five other protein sequences from different organisms obtained from BLAST searches. These sequences included "Chain S, Green fluorescent protein [Escherichia coli]", "Chain A, NODAMURA

VIRUS COAT PROTEINS [Nodamura virus]", "Chain S, Green fluorescent protein [Aequorea victoria]", "Chain A, NPH1-1 [Avena sativa]" and finally "Chain A, Lipoprotein [Burkholderia cenocepacia J2315]." Each sequence was chosen based on its E-value, a measure of sequence similarity, with varying degrees of relatedness to a target protein of interest, this can be seen here [F1].

An Insite of possible coincidence obtained from this phylogenetic tree [F1], is that the Chain S of Green fluorescent protein found in the organism Aequorea Victoria ("crystal jelly"), appears to have the exact same property as the Chain S of the Green fluorescent protein from our Escherichia coli organism. This may indicate a true linkage in homology between the two proteins and can show how the same protein found in a bacterium is also used in the crystal jelly fish that provides the property of fluorescence in a marine environment.

As seen in my phylogenetic tree, Avena Sativa (oats) stands out from the rest, there are a few reasons for this. It's possible that the sequence in Avena sativa and C3UPD2_ECOLI protein chain are homologous, sharing a common ancestry but the protein will most likely have different purposes within their own protein complex. But it could also be that the data produced in the laboratory about Avena Sativa was contaminated with our protein strain which could have led unexpected results such as this, which is unlikely but clarification must be done on the matter. This could instead by a false positive from BLAST, where a hit is found due to a coincidence in similarity and the chain that was searched for just happened to be the same and could even have a different purpose in the overall strain it resides in. But it was left in the phylogenetic tree due to the possibility of possible homology that can be further researched to clarify in the future.

Furthermore, after further research, the first Green fluorescent protein was even first isolated and characterised from the jellyfish species Aequorea Victoria and the natural role of GFP in Aequorea Victoria is to produce bioluminescence but not fluorescence.

## Methods and tools used to build a 3D structure model

Due to the lack of data on this C3UPD2 · C3UPD2_ECOLI, there is no individual models of the protein by itself on PDB, UniProt, SWISS-MODELrather, AlphaFold, NCBI and Expacy. So I had to resort to searching a protein complexs that included the amino acid structure "ANDENYALAA" in any one of its strains.

This led me to choose the sequence "ANDENYALAA" associated with the PDB accession of "8ET3_S", Chain S, Green fluorescent protein [Escherichia coli], which is the same protein complex stated before. The data about this structure was obtained through electron microscopy and was added to the database on October 16th 2022, and released on August 9th 2023. This suggests how new and possibly even how incomplete the information about this protein actually is and must be taken into consideration, and the model of "ANDENYALAA" can be seen here [F2].

Due to the fact that C3UPD2 · C3UPD2_ECOLI has only 10 residues, sites like PSIPRED, LOMETS, Zhang's I-Tasser would NOT accept a chain of such short length, all of them only accept residue lengths longer than 20. Because of this I had to resort to only using the simpler method of modelling five models of the protein and aligning it to a model that was found within the Green Florescent Protein as seen here [F3].

So the generation of our own possible PDB models was needed with the use of modeller with the target chain of "ANDENYALAA". Modeller then provided five target models with various DOPE scores in which Target_1 was chosen due to being the lowest DOPE score, as seen here [F4]. This was then modelled beside the PDB model of the chain found in the GFP complex. The models were then

aligned and a sticking resemblance was seen between the two, which was expected with such small sequences with 100% sequence alignments and identical lengths.

## Explanation of how you identified the protein's network of interactions

C3UPD2 · C3UPD2_ECOLI is located in Chain S of the Green fluorescent protein (GFP) within Escherichia coli which is a unicellular bacterial organism. But where this protein is specifically located within the cell of E.coli is a different question, because it is commonly used as a marker in molecular biology and cell biology studies. So when the GFP is used as a reporter or a fusion protein, its specific location within the cell depends on how it is expressed and tagged within the bacterium.

GFP can be located in many different cellular compartments depending on the whether it was developed naturally from natural selection, especially within aquatic circumstances like in Aequorea victoria or from human experimental setups.

GFP has been cloned and adapted by researchers for use as a marker in a variety of organisms, including bacteria, yeast, plants, and animals [R2]. And in our instance it is most probable that GFP is NOT found naturally in E.coli [R3]. Scientists have introduced the gene encoding GFP into E. coli and other organisms to study gene expression, protein localisation, and various other cellular processes in E. coli and these BLAST results do not mean that GFP is naturally occurring in E. coli but instead, they just suggest that there is a sequence similarity between the two.

And thus, the green florescent protein complex which where our C3UPD2 protein strand is found in, is most likely found in the cytoplasm of the E. coli cell. This is achieved by genetically engineering the E. coli to produce GFP, where the gene is inserted into the E. coli's DNA which is when the bacteria will then produce GFP on its own, causing it to be present in the cytoplasm.

## Evaluation of the UniProt record's accuracy

Now UniProt has some concise and accurate data on C3UPD2 · C3UPD2_ECOLI, starting with the first FASTA data being added to the database back in 2009, but has only had 2 publication relating to the protein ever since and one of which being unnamed. It still lacks in a lot of essential information such as the lack of data on experiment usage of this exact protein which can lead to UniProt having an uninformed idea of the function, interactions and the structure of the protein. Following that, the 3D model wasn’t uploaded to its catalogue which I could add from my own extraction of it from the GFP complex. Additionally there is some discrepancies such as the information on whether or not the protein is obsolete within the protein complex, details on how could it help in the transfer of energy within the GFP to create green florescent light and if so how. But if not, any details on where this very short strand of amino acids came from, and what’s it’s exact homology?

Conclusively, the accuracy of UniProt is rightfully marked at a notation level of 1/5. This protein needs further research, experimental usage and homology data to have a concise view of C3UPD2 · C3UPD2_ECOLI and how it impacts the Green Florescent Protein. This might be due to the fact that the individual protein itself doesn’t do anything special or anything of significance but is rather a simple cog with the whole mechanism of the GFP.

## Summarise the main findings and their significance

Overall the findings from this research on C3UPD2 · C3UPD2_ECOLI is that there is a scarcity of information and underrepresentation in public databases. The homology and Sequence Analysis showed 100% sequence identity with multiple other sequences from a range of organisms allowing to map a rough idea of possible homologies of the protein. A strong relationship with the GFP found in many organisms including E.coli which was found to be commonly used as molecular and cellular

marking for various studies. Leading to the Phylogenetic analysis revealing its lineage of the GFP with Aequorea Victoria that showed a shared role in providing fluorescence properties or a cell. The location of C3UPD2 · C3UPD2_ECOLI within the GFP complex and the most likely location of the protein within the cell believed to be in the cytoplasm of the E.coli cell in this case. And finally Uniprot's records for C3UPD2 · C3UPD2_ECOLI having a lack of crucial information, especially about experimental implementation data and the its homology.

## List of references for all sources and tools used

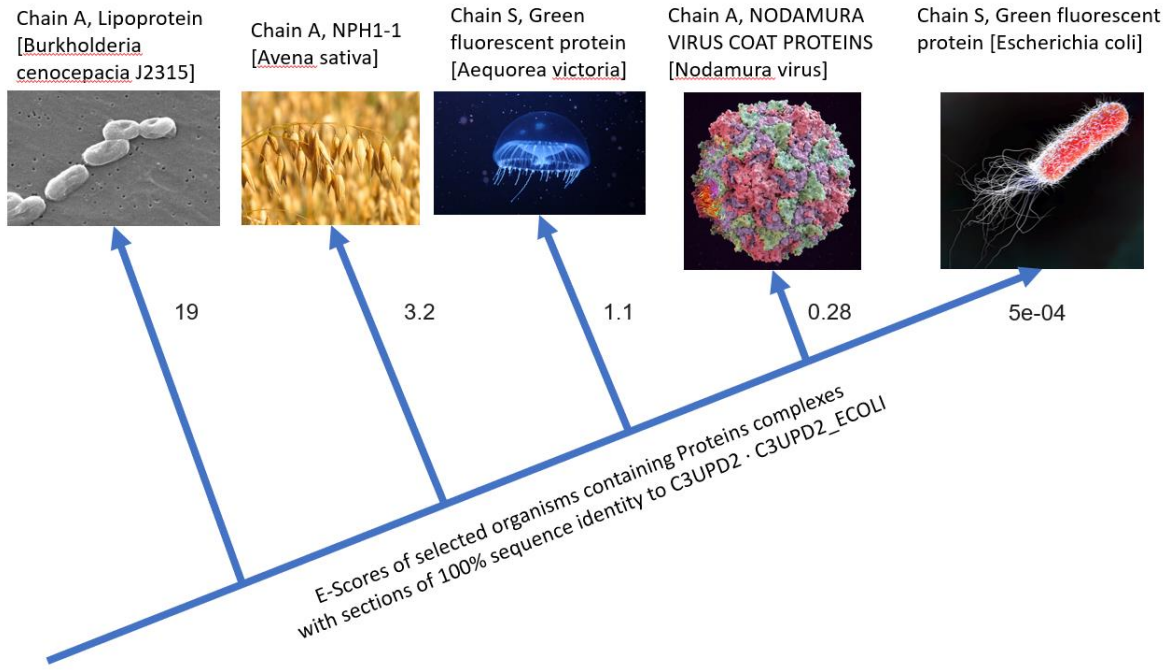
[R1] Protein Secondary Structure, Alpha Helices, Beta Sheets, Hairpins And Loops (no date b). https://proteinstructures.com/structure/secondary-structure/

[R2] Soboleski, M.R., Oaks, J.L. and Halford, W.P. (2005) 'Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells,' The FASEB Journal, 19(3), pp. 1–20. https://doi.org/10.1096/fj.04-3180fje

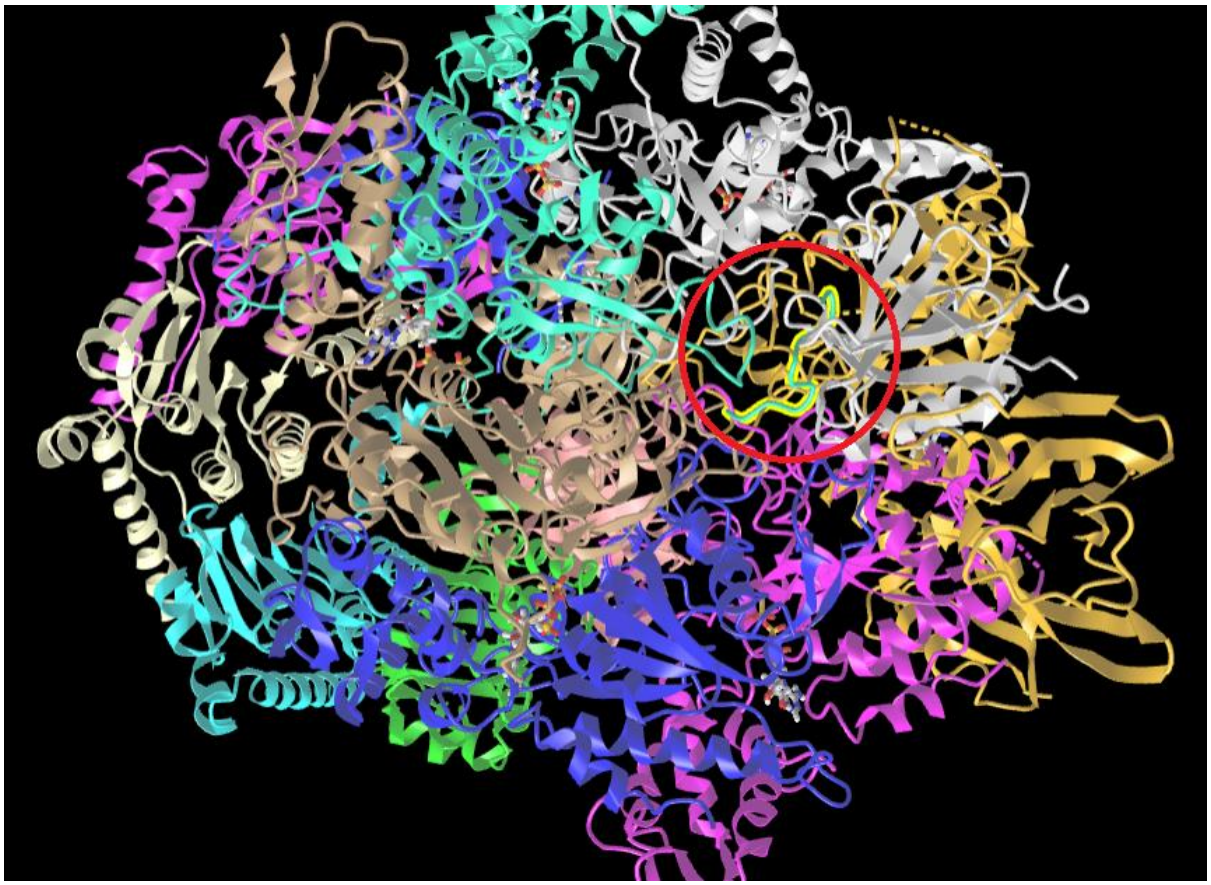
[R3] Joon, H. et al. (1999) 'Green Fluorescent Protein as a Noninvasive Stress Probe in Resting Escherichia coli Cells,' Applied and Environmental Microbiology, 65(2), pp. 409–414. https://doi.org/10.1128/aem.65.2.409-414.1999

## Figures and tables

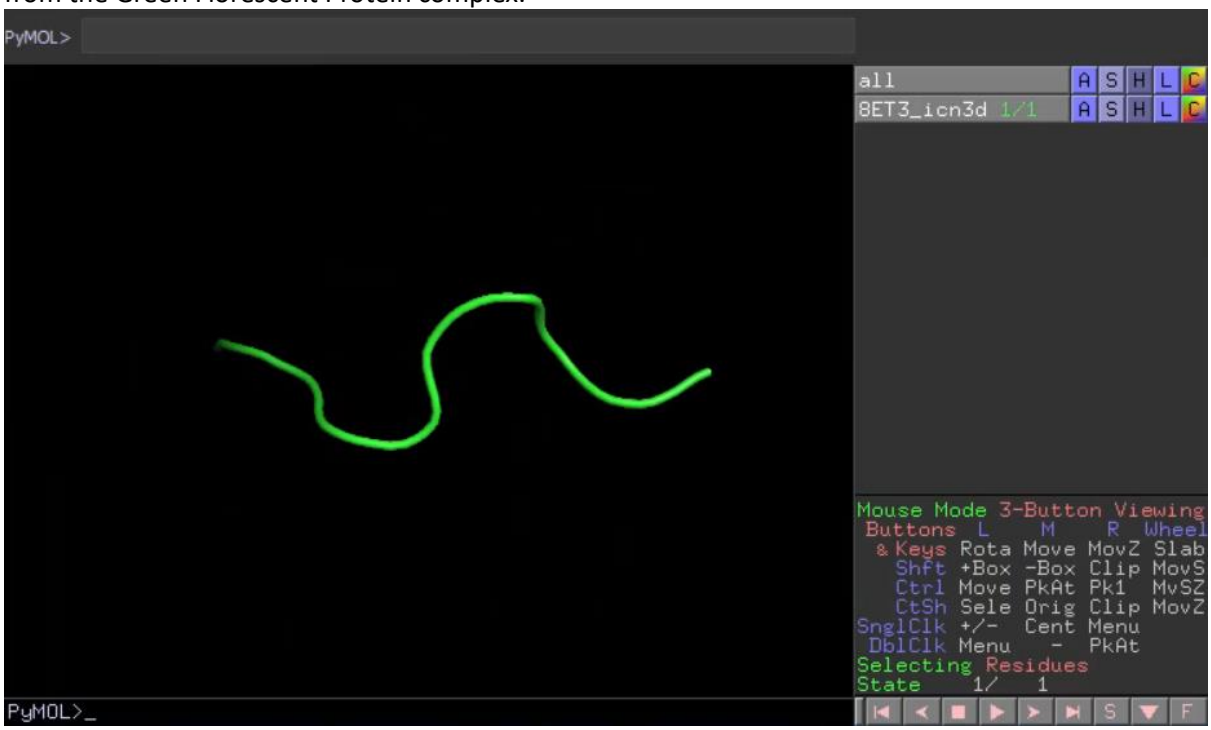
[F1] phylogenetic tree of the GFP found in E.coli



[F2] The independent sequence “ANDENYALAA” found in “Chain S, Green fluorescent protein [Escherichia coli]”



[F3] The structure of “ANDENYALAA” (C3UPD2 · C3UPD2_ECOLI) named as “8ET3_icn3d”, extracted from the Green Florescent Protein complex.

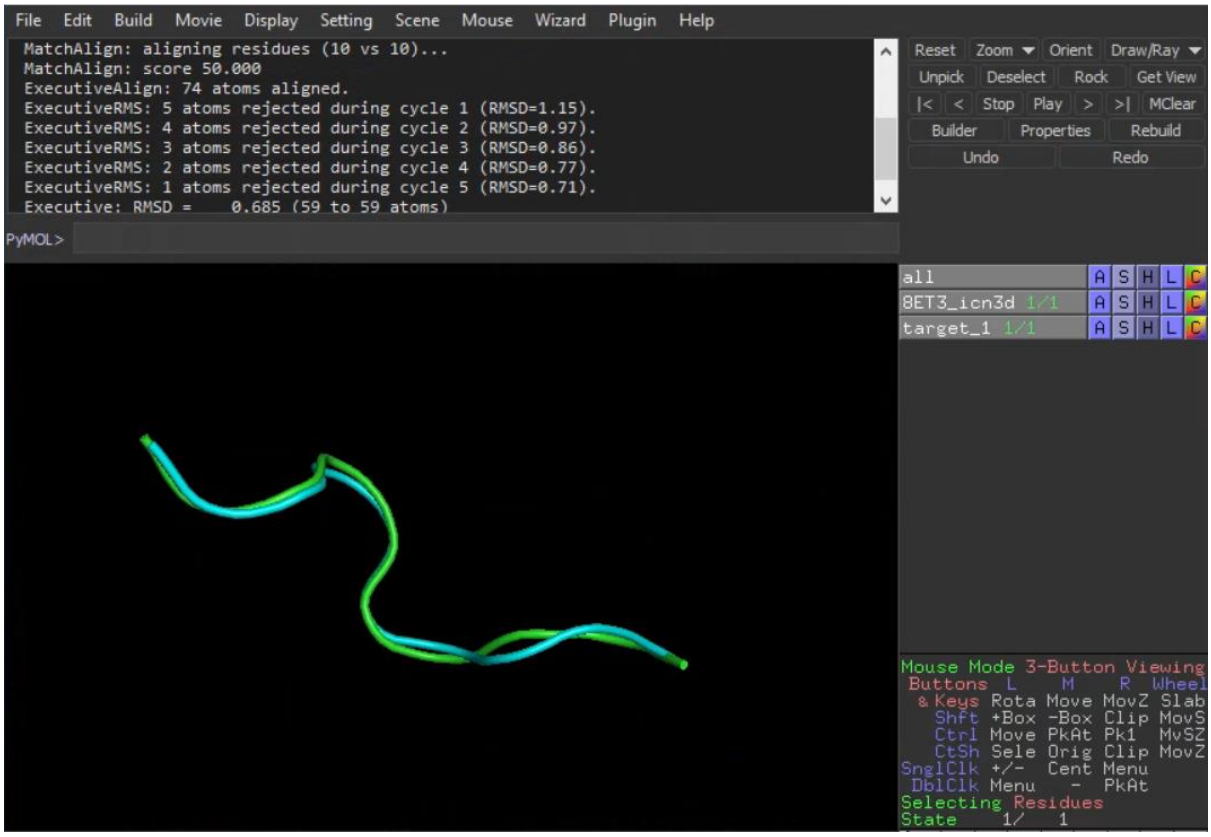


[F4] Selection of generated models, Target_1 was chosen as it has the lowest DOPE score.

```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
target_1.pdb                    24.04699     -299.25293        0.53861
target_2.pdb                    30.14864     -263.23795        0.84864
target_3.pdb                    28.46985     -270.28293        0.80810
target_4.pdb                    36.13095     -267.59738        0.64708
target_5.pdb                    31.14869     -289.82681        0.48653


Dynamically allocated memory at          finish [B,KiB,MiB]:      5774552    5639.211     5.507
Starting time                                                : 2023/10/24 19:32:25
Closing time                                                 : 2023/10/24 19:32:36
Total CPU time [seconds]                                     :       3.09
```

[F5] The generated model “Target_1” and the “8ET3_icn3d” model extracted from the Green Florescent Protein Complex




C3UPD2 · C3UPD2_ECOLI FASTA data:
>tr|C3UPD2|C3UPD2_ECOLI TmRNA tag peptide OS=Escherichia coli (strain K12) OX=83333 PE=4 SV=1
ANDENYALAA