Integrative functional and structural characterization of uncharacterized protein

Author: Payal Jain, Mohit Fulara
Date: 12 Feb 2025

ABSTRACT

This study focuses on the functional and structural characterization of the uncharacterized protein **A0A0H2Z666 (yjgQ)** from *Escherichia coli* O1:K1 using computational bioinformatics tools. The protein sequence was analyzed using **BLASTp**, which identified a high-confidence match with **LptG, a lipopolysaccharide export system permease protein**, with **99.72% identity and 100% query coverage**. Functional annotation through **InterProScan and QuickGO** revealed its role in **membrane-associated transport and lipid transfer processes**. Structural modeling using **Swiss-Model** provided a **high-quality 3D structure**, which was validated by **Ramachandran plot analysis**. Binding site predictions using **COACH-D and F-pocket** identified potential ligand-binding pockets. These findings suggest that **A0A0H2Z666 (yjgQ) is a likely LPS transport-associated permease protein**, playing a key role in **cell envelope maintenance in E. coli**. Future experimental validation is required to confirm these predictions.

1. INTRODUCTION

Uncharacterized proteins,also known as hypothetical proteins, are proteins whose biological functions remain unknown or poorly annotated despite the availability of their amino acid sequence. These proteins of unknown function are a barrier to our understanding of molecular biology. Assigning function to an uncharacterized proteins is crucial for advancing study in molecular biology.Functional annotation of unknown proteins can aid in identifying potential drug targets.
Despite being most-widely studied organisms,significant portion of E.Coli genes still lacks functional annotation. Predicting the functions of these uncharacterized proteins can help elucidate their biological roles within the cell, providing a more complete "systems-wide" functional blueprint of the microbe. Addressing the gaps in the functional annotation of genes enhances our understanding of basic prokaryotic traits and evolutionary significance.
This study aims to annotate the selected uncharacterized protein to predict its function ,structure and its biological interactions using computational tools.

In this study, we have selected **yjgQ gene** having an **Open Reading Frame (ORF) APECO1_2132** that encodes the uncharacterized protein **A0A0H2Z666**, in ***Escherichia coli* O1:K1 / APEC** . This selected protein is **361** amino acids long,with a molecular weight of 39,720 Da.
To predict and validate the functions of uncharachterized proteins,we employed bioinformatics based analysis such as:-
Sequence alignment (BLASTp) to find the homologous proteins.
Domain search (InterProScan) to predict conserved functional domains.
Structure modelling and visualization (Swiss-Model,I-Tasser,PyMol) to analyze secondary and tertiary structures.
Pathway mapping (KEGG,STRING, Reactome) to identify potential biological interactions.
By this computational analysis we aim to identify functional domains,structural motifs,active sites,binding pockets, ligand-binding sites and biological pathway associations, providing the insights into the potential role of A0A0H2Z666 in *E.Coli* cell.

2. METHODOLOGY:

2.1 Protein Selection:

- The uncharacterized Protein **A0A0H2Z666** retrieved from the UniprotKB database(TrEMBL) in FASTA format.
- Gene Name: yjgQ

- Accession ID : A0A0H2Z666
- Source Organism: *Escherichia coli* O1:K1
- Amino acids: 361
- Status: Unreviewed (TrEMBL)

2.2 Sequence Annotation Analysis

2.2.1 Sequence alignment

To find the homologous proteins , BLASTp of NCBI was used.

- The fasta sequence of query protein was uploaded.
- Expect threshold: 0.01
- Scoring matrix: BLOSUM62
- Database used: UniprotKB/Swiss-Prot
- Other search parameters were kept to default settings.

2.2.2 Domain Search and GO Term analysis

- InterProScan was used to identify the functional domains,families and structural motifs.
- Domains are the structural or functional unit of protein each contributing to overall activity of the protein.
- The sequence was submitted to the database .
- Gene Ontology (GO) terms were extracted using InterPro and validated using QuickGO.
- Uniprot ID was entered into QuickGO, and the associated GO terms were compared with InterPro results.

2.3 Structural Analysis

2.3.1 Secondary Structure Prediction

- PSI-PRED 4.0 tool was used to predict the secondary structure elements of selected protein.
- PSI-PRED utilizes protein-specific scoring matrices (PSSM) generated from the sequence to predict the alpha-helices, beta sheets,coil regions, transmembrane matrix,etc.
- Fasta file was uploaded in the database

2.3.2 Tertiary Structure Prediction

- SWISS-MODEL was used for homology-modelling of protein's tertiary structure.
- FASTA sequence was uploaded and templates were searched.
- Template (PDB ID: C3SEX2.1.A) was selected on basis of :
    - Sequence identity 99.72%
    - Similarity of 60%
    - High sequence coverage
- The 3D structure was predicted and its quality was validated using Ramachandran plot,which checks the dihedral angle distribution of the protein structure.
- The modeled protein was downloaded in PDB format.

2.4 Structural Visualization

2.4.1 Binding sites Prediction:

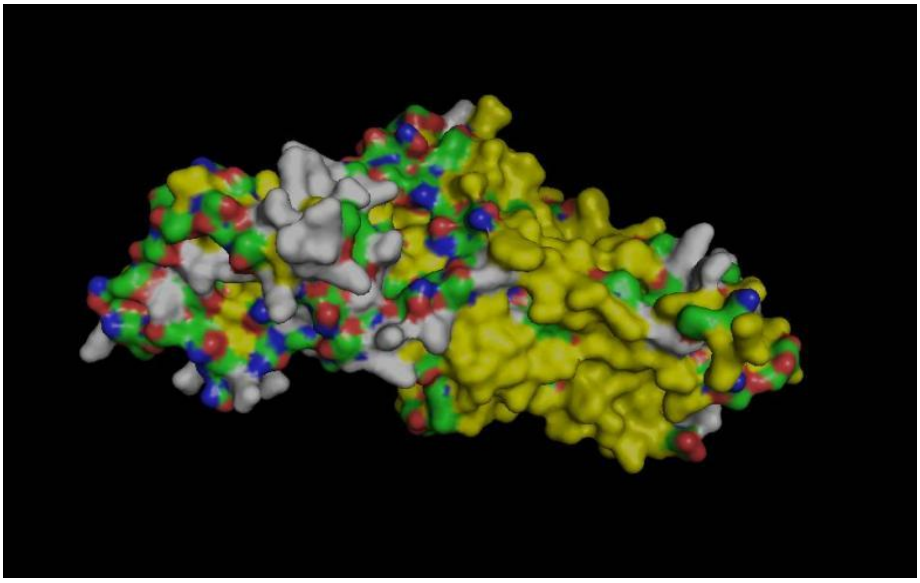
- COACH-D was used for predicting binding sites, active sites of the modeled protein.
- Modeled PDB file was submitted as input in COACH-D web server.
- COACH-D combines multiple binding site prediction tools such as COFACTOR, TM-SITE, S-SITE.
- The results includes:
    - **Predicted binding residues** and their positions.
    - **Template-based ligand-binding information**.
    - **Cluster analysis for binding site conservation**.
- F-pocket server was used to detect the cavities in a protein's 3D structure that may serves as potential ligand-binding sites.PDB file of modeled protein was uploaded.
- The detected pockets were scored and ranked based on volume, druggability potential, and alpha sphere representation. The detected pockets were mapped onto the protein structure and color-coded for better interpretation.

2.4.2 Visualization under PyMOL

- The PDB formatted modeled protein was loaded in PyMOL for structure visualization.
- The predicted binding pockets from F-pocket and COACH-D server were mapped onto the protein structure.
- The protein was rendered in surface representation to examine the overall topology and accessibility of binding pockets.
- The protein was displayed in cartoon representation, with binding sites highlighted in red stick representation to enhance visibility.

RESULTS AND DISCUSSION:

1. BLASTp Homolgy Result:



The BLASTp search was conducted to identify homologous sequences for the uncharacterized protein A0A0H2Z666(yjgQ) from E.coli. The analysis revealed a high confidence match with lipopolysaccharide export system permease protein LptG from Escherichia coli K-12. This best blast hit exhibited:

- 99.72% sequence identity
- 100% query coverage
- Highly significant E-Value of 0.0

- The second match was found with the LptG protein from Haemophilus influenzae Rd KW20 showing 53.82% identity and 98% query coverage,but with a significantly higher E-value and lower score,indicating weaker homolog.

Protein	Accession	Organism	Accession Length	Max Score	Query Cover	E Value	Percent Identity
LptG (Lipopolysaccharide export system permease protein)	P0ADC6.1	Escherichia coli K-12	360	663	100%	0	99.72%
LptG (Lipopolysaccharide export system permease protein)	P45332.1	Haemophilus influenzae Rd KW20	358	363	98%	1e-123	53.82%

Interpretation:

BLASTp results confirm that A0A0H2Z666 (yjgQ) is highly homologous to **LptG from E.coli K-12** ,which predicts that it functions as part of lipopolysaccharide export system, which is essential for bacterial outer membrane synthesis. **yjgQ** is likely an **LPS transport-associated permease protein** ,contributing to cell envelope maintenance in E.coli O1:K1.

2. **Domain and Functional Annotation using InterPro**

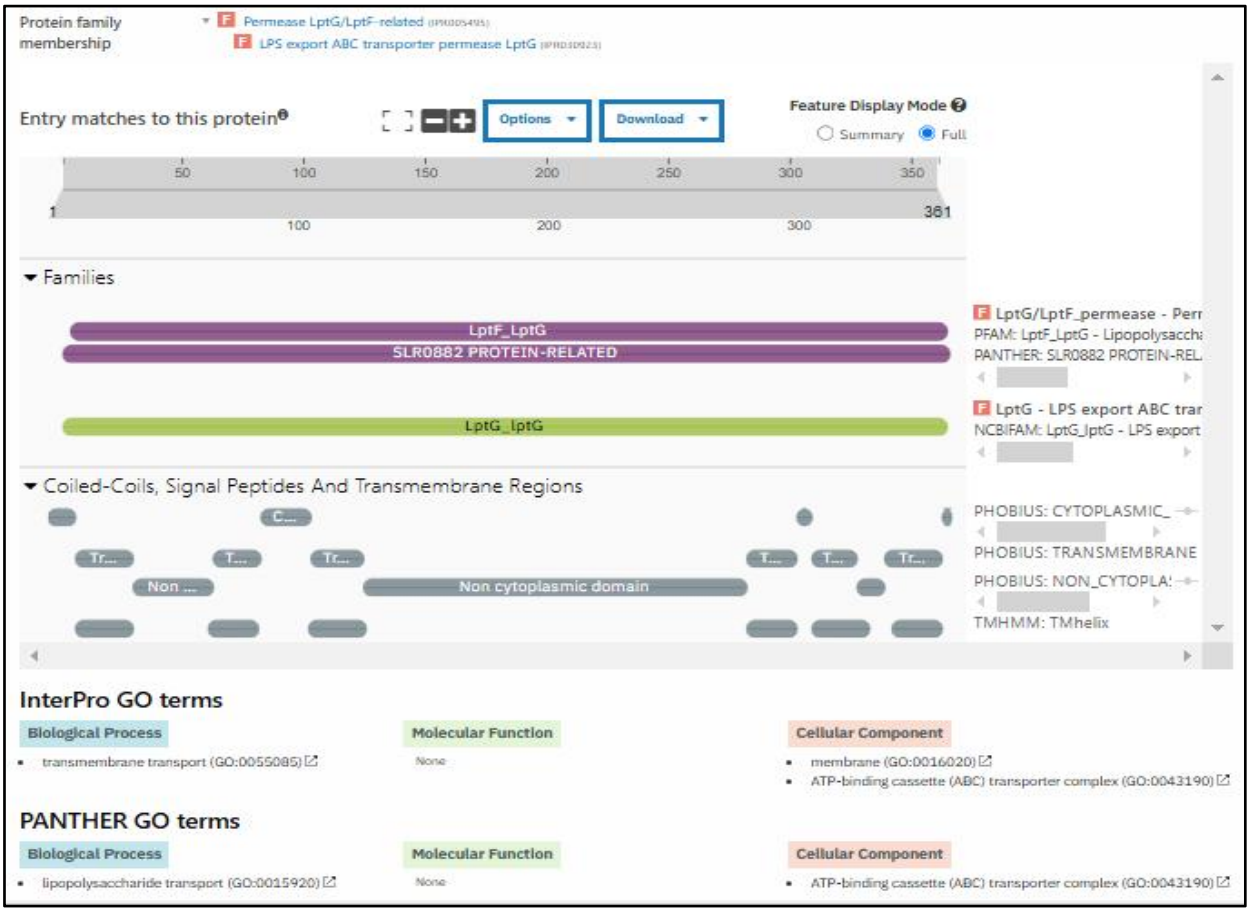
- The InterProScan results revealed that the protein belongs to the **LptE_LptE family**, specifically involved in **LPS-assembly protein LptE**. The following features were identified:

  - **Protein Family:** LptE_LptE superfamily
  - **Identified Domain:** LptE_LptE domain (functional role in membrane-associated transport)
  - **Structural Features:**
    - Presence of **coiled-coil regions**
    - Identification of **signal peptides and transmembrane regions**
    - Predicted **non-cytoplasmic domain**, indicating its extracellular or periplasmic localization

- Additionally, **Gene Ontology (GO) terms** were mapped:

  - **Biological Process:** Lipid transport
  - **Molecular Function:** ATP-binding cassette transporter activity
  - **Cellular Component:** Membrane-associated

These findings suggest that the protein may function in **membrane-associated transport** and lipid transfer processes.



Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With/From
UniProtKB:A0AHZ2666	yjgO	involved_in	GO:0015920 (Lipopolysaccharide transport)	ECO:0007826	GO_REF:0000118	PANTHER:PTN002128460
UniProtKB:A0AHZ2666	yjgO	involved_in	GO:0055085 (Transmembrane transport)	ECO:0000256	GO_REF:0000022	InterPro:IPR000923
UniProtKB:A0AHZ2666	yjgO	located_in	GO:0005886 (Plasma membrane)	ECO:0000256	GO_REF:0000120	UniProtKB:KW-1003
UniProtKB:A0AHZ2666	yjgO	located_in	GO:0016020 (Membrane)	ECO:0000256	GO_REF:0000022	InterPro:IPR003439
UniProtKB:A0AHZ2666	yjgO	part_of	GO:0043190 (ATP-binding cassette (ABC) transporter complex)	ECO:00005 1	GO_REF:0000120	InterPro:IPR000923

- **QuickGO analysis**

QuickGO analysis identified GO:0005886 (Plasma membrane) as a potential localization for the yjgQ protein. This term was not retrieved from InterPro, indicating that QuickGO can provide complementary insights into the protein's functional role. The annotation suggests a membrane-associated function, which aligns with its predicted role in lipid transport.

3. Secondary Structure Prediction Using PSIPRED

The secondary structure prediction was done by PSIPRED on the target protein sequence. The results show α-helices, β-strands, and coil regions, as indicated by their PSIPRED output.

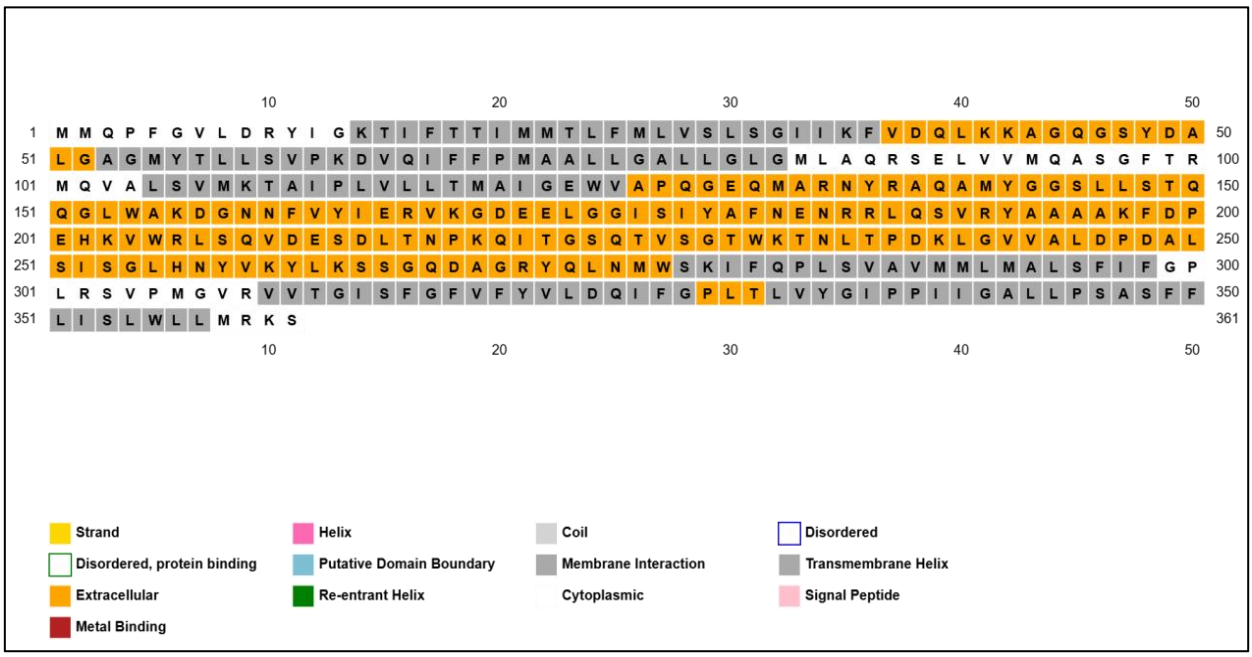
- α-Helices:

PSIPRED predicts a large fraction of the protein sequence, comprising α-helices. The regions are depicted in pink and suggest a helical conformation that may impart stability and function to the protein.

- β-Strands:

The presence of β-strands, as shown in yellow, indicates that these regions may be involved in the formation of β-sheets. These regions are quite important for forming the structural core of a protein, in which inserts would lead to the formation of a β-barrel or β-sheet in globular proteins.

- Coil Regions:
  Coil regions (represented in black) indicate flexible, unstructured regions within the protein. These regions may contribute to protein dynamics, loop formation, or interaction sites.
- Confidence Scores:
  The confidence scores for each predicted secondary structure are represented in blue bars, with higher values indicating stronger confidence in the assigned structure. The confidence levels vary across the sequence, with some regions showing strong prediction reliability.



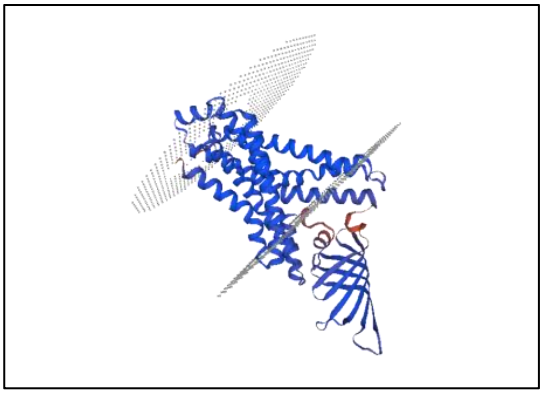
Interpretation

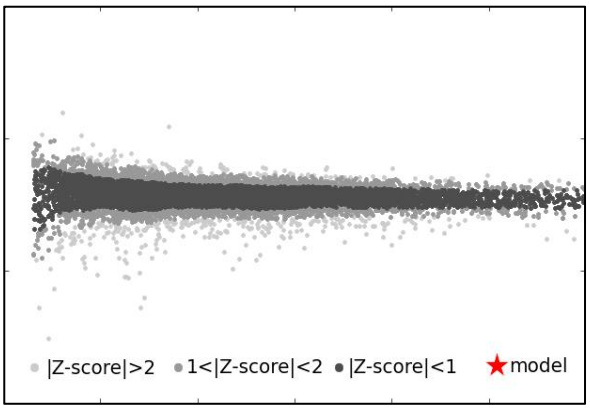
A significant portion of helices present may be indications of a helical fold that are widely known to occur in transmembrane or globular proteins. The strands suggest the possibility of the formation of sheets acting as stabilizers.The regions of coil also imply possible flexible regions that can contribute to their functionality.The

predicted secondary structures thus give insight into the folding pattern and structural organization to be verified further by homology modeling.

4. Structure Modelling

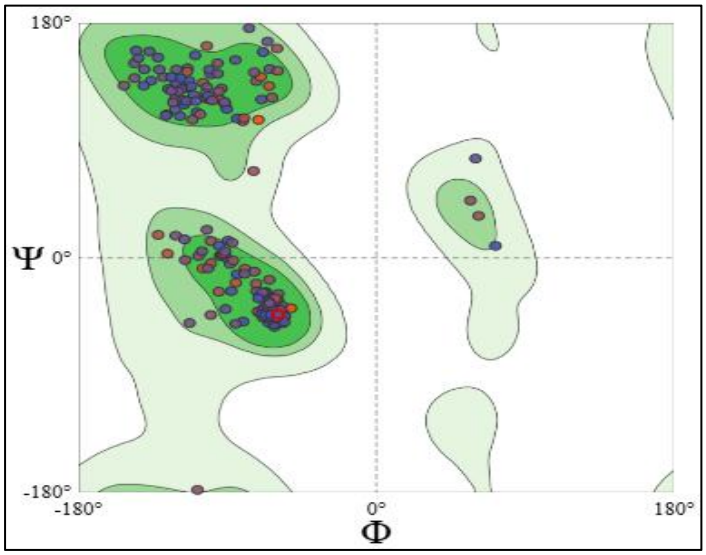
- FASTA sequence of unchararcterized protein was uploaded in SWISS-MODEL and potential template was selected.
- The 3D structure of the target protein was successfully predicted using Swiss-Model, generating a homology-based model.





- The QMEAN Z-score was analyzed to evaluate the overall quality of the model. The QMEAN score distribution indicates that the predicted model aligns well with experimentally determined protein structures, suggesting a reasonable model quality.

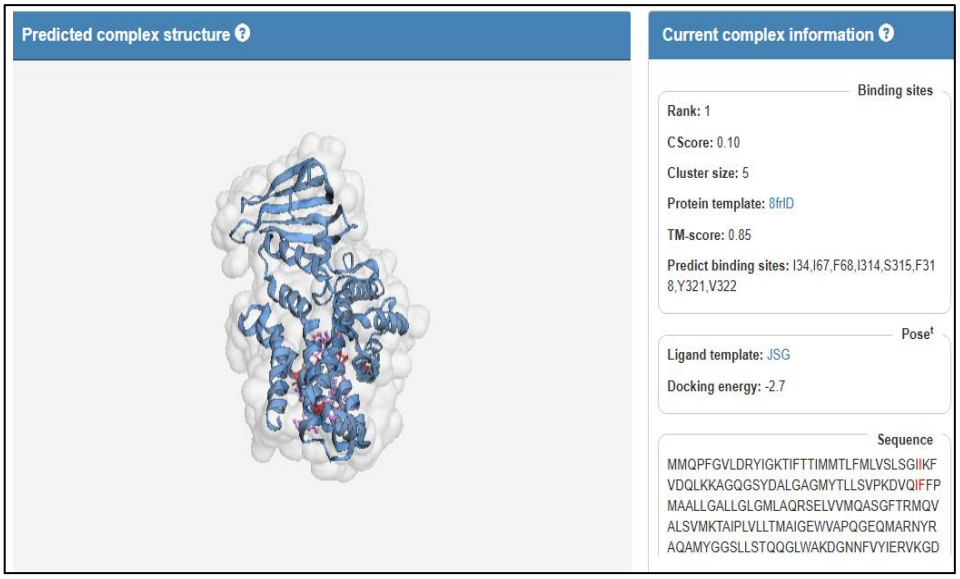
- **Ramachandran Plot Analysis:-** The Ramachandran plot (Figure X) was generated to assess the stereochemical quality of the predicted structure. The results indicate that:
  - 98.89% of residues fall in favored regions,
  - 0% are in outlier regions,
  - 0.67% of residues are classified as rotamer outliers.



These statistics confirm that the backbone dihedral angles (Φ, Ψ) of most residues are within energetically favorable regions, indicating a highly reliable structural model.

5. Structural Visualization:

- The **COACH-D** analysis provided insights into the potential active and binding sites of the protein model.The best-ranked binding site prediction was associated with the template 8frID. The key binding residues identified were I34, I67, F68, I314, S315, F318, Y321, V322. These residues are likely involved in ligand interactions and could serve as functional hotspots.

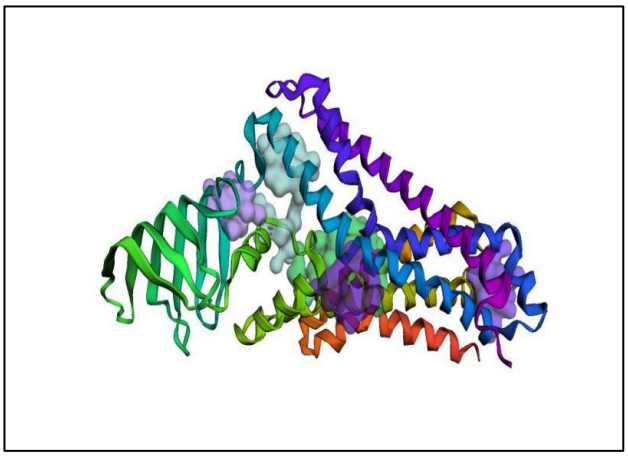


- **COACH-D** predicted the most relevant ligand for binding as JSG.Another ligand (L0W) was detected with a lower confidence score in template 6s8nD.
- The best-ranked binding cluster had a C-Score of 0.10, indicating a moderate confidence level. The TM-score for the best binding template was 0.85, showing strong structural similarity.
- The **F-pocket** analysis identified 26 potential ligand-binding pockets in the modeled protein structure. These pockets were ranked based on their score, druggability, volume, and number of alpha spheres.The best 5 ranked pockets are listed below:

Pocket	Colour	Score	Druggability	Alpha Spheres	Volume
1.	Purple	0.142	0.001	21	101.955
2.	Green	0.131	0.001	124	1061.676
3.	Dark Purple	0.118	0.051	31	320.281
4.	Blue	0.057	0.025	38	301.985
5.	Cyan	0.048	0.006	50	508.444

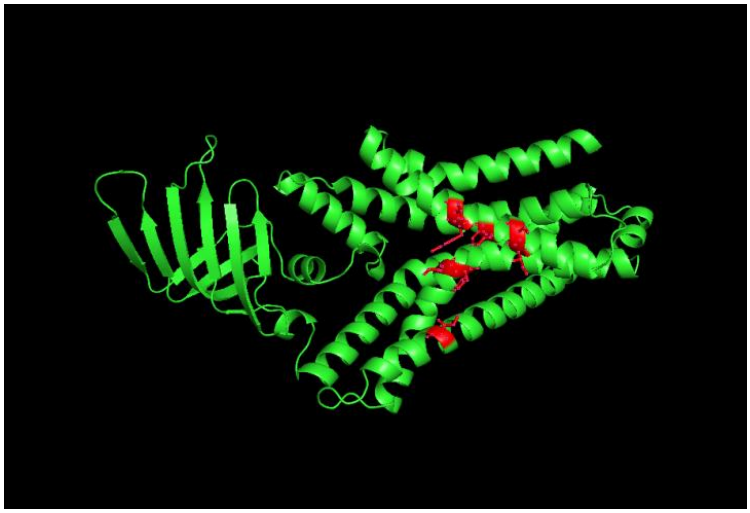
- **Pocket 1** had the highest score and a moderate volume
- **Pocket 2** had the largest volume and a high number of alpha spheres indicating highly accessible cavity
- **Pocket 3** had a higher druggability score compared to others , suggesting potential for ligand interaction.
- **Pocket 4 and 5** had moderate scores and lower druggability values, making them less favorable binding sites.

The 3D visualization of the protein structure with detected pockets is shown in Figure :



6. Visualization under PyMOL

- **Binding site prediction from F-Pocket and COACH-D identified ligated-accessible regions**, which were mapped onto the protein structure.
- In **cartoon representation**, the binding sites were predominantly located in **helical and loop regions**, indicating potential flexibility and interaction points.
- The **surface representation showed distinct pocket formations**, confirming ligand accessibility.
- These results provide a detailed structural insight into the binding regions, aiding in ligand docking and drug design applications.



CONCLUSION:

This study provides strong computational evidence that the uncharacterized protein A0A0H2Z666 (yjgQ) functions as an LPS transport-associated permease protein in Escherichia coli O1:K1. The BLASTp results, coupled with domain identification (InterProScan) and structural modeling (Swiss-Model), support this functional prediction. The presence of membrane-associated transport domains and ATP-binding motifs further reinforce its potential role in bacterial outer membrane biogenesis. Binding site analyses suggest possible ligand interactions, opening avenues for drug targeting and antibiotic resistance studies. Future work should focus on experimental validation through functional assays and in vitro binding studies to confirm its role in LPS transport. The integration of computational and experimental approaches will be crucial in refining our understanding of this protein's biological function.

REFERENCES:

1. BLASTp:
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
2. InterProScan Database: https://www.ebi.ac.uk/interpro/search/sequence/
3. PSI-PRED Database : http://bioinf.cs.ucl.ac.uk/psipred/
4. Swiss-Model: https://swissmodel.expasy.org/
5. F-pocket: https://durrantlab.pitt.edu/fpocketweb/
6. COACH-D: https://yanglab.qd.sdu.edu.cn/COACH-D/