CAPSTONE PROJECT

TITLE

In-silico Identification and Functional Characterization of a Hypothetical Protein (A0A8S4BTT0) from Menidia menidia Using Sequence Analysis and Homology-Based Annotation

Submitted by

Swathi PS
BS Chemistry
SRMIST

Under the Guidance of

Ms. Muskan Kashyap
Bioinformatics Analyst
Founder, Biotechshalaa – YouTube Channel

Project Duration
25.01.2026 – 05.02.2026

ACKNOWLEDGEMENT

I sincerely thank **Ms. Muskan Kashyap**, Bioinformatics Analyst and Founder of *Biotechshalaa (YouTube Channel)*, for her valuable guidance, continuous support, and insightful suggestions throughout the course of this project. Her expertise and encouragement were instrumental in the successful completion of this work.

1. Abstract

Hypothetical proteins constitute a significant fraction of sequenced genomes, yet their biological roles remain unclear. In this study, an uncharacterized protein from *Menidia menidia* (UniProt ID: A0A8S4BTT0) was analysed using an in-silico homology-based approach. Sequence analysis and BLASTP search against the Swiss-Prot database revealed significant similarity to the MHC class II invariant chain (CD74). Based on conserved regions, alignment quality, and functional annotation of homologs, a putative function was assigned to the target protein. This study demonstrates the utility of computational methods in functional prediction of hypothetical proteins.

2.Keywords

Hypothetical protein, BLASTP, Homology analysis, Functional annotation, *Menidia menidia*, BioPython

3.Introduction

Hypothetical proteins constitute a significant portion of sequenced genomes, yet their biological functions remain unknown due to the lack of experimental characterization. Understanding the roles of these proteins is essential for gaining insights into cellular processes and organismal biology. In-silico approaches, particularly homology-based analysis using sequence alignment tools, provide an efficient and cost-effective method for predicting protein function. In the present study, a hypothetical protein from *Menidia menidia* was analysed using BioPython-based BLAST and sequence analysis tools to infer its potential function.

4.Methodology

4.1 Sequence Retrieval

The amino acid sequence of the selected hypothetical protein from *Menidia menidia* was retrieved from the UniProt database using its unique UniProt accession ID. The sequence was downloaded in FASTA format and used as the input for all subsequent in-silico analyses.

4.2 Sequence Quality Analysis

The retrieved protein sequence was examined for completeness, length, and the presence of ambiguous amino acids. Basic sequence properties such as sequence length and amino acid composition were assessed to ensure the sequence was suitable for downstream bioinformatics analysis.

4.3 Sequence Filtering and Validation

The sequence was validated to confirm that it represented a true protein-coding sequence and not a fragment or low-quality prediction. Redundant or truncated sequences were excluded, and only the validated full-length sequence was considered for homology analysis.

4.4 Homology Search (BLAST Analysis)

STEP-1

A homology search was performed using the BLASTp program through the BioPython interface against the Swiss-Prot protein database. Significant hits were identified based on E-value, alignment score, and sequence similarity. The top-ranking hits were analysed in detail to determine evolutionary conservation and functional relevance.

STEP-2

The BLASTP search of the hypothetical protein sequence (A0A8S4BTT0) against the Swiss-Prot database yielded a total of 23 hits. This indicates multiple homologous sequences exist in the database, which allows for comparative analysis to infer potential function. All BLAST hits were listed in order of similarity, ranked by E-value and alignment score. The top hits were closely related class II histocompatibility antigen gamma chains from Rattus norvegicus, Mus musculus, and Homo sapiens, suggesting conserved functional regions.

STEP-3

The top three hits were evaluated based on alignment score and E-value. The best hit was identified as sp|P10247.2| H-2 class II histocompatibility antigen gamma chain (Rattus norvegicus) with an alignment score of 333.0 and a highly significant E-value of 8.56×10^{-36}. This suggests strong homology between the query and this protein.

STEP-4

The highest-scoring segment pair (HSP) from the closest hit was selected to examine the most significant alignment. The HSP had a score of 333.0 and an E-value of 8.56 × 10^-36, confirming a highly reliable alignment region for further analysis.

STEP-5

The alignment between the query and subject sequences spans residues 24–264 of the query and 21–255 of the subject. The alignment shows multiple conservative substitutions, indicating that key functional residues are preserved. Gaps are present but minimal, suggesting overall structural and functional conservation.

- Query sequence: shows residues of hypothetical protein
- Matched sequence: shows residues in the known protein
- Alignment match: "+" indicates conservative substitutions

Conservative substitutions indicate potential preservation of biochemical properties and likely retention of similar function.

STEP-6

The mapped functional region of the query corresponds to residues 24–264, aligning with residues 21–255 of the reference protein. This region likely contains the functional domains characteristic of class II histocompatibility gamma chains, such as the invariant chain domain involved in antigen presentation.

4.5 Functional Annotation

Functional annotation was inferred by comparing the hypothetical protein with well-characterized homologous proteins obtained from BLAST results. Conserved regions, alignment coverage, and functional domains present in the matched sequences were used to predict the probable biological function of the hypothetical protein.

5.Result

The BLAST analysis identified CD74 (H-2 class II histocompatibility antigen gamma chain) as the closest homolog of the hypothetical protein A0A8S4BTT0. CD74 is a key protein in the immune system, functioning as the invariant chain that stabilizes MHC class II molecules and guides antigenic peptides for presentation to T-cells. Alignment of the query protein with CD74 revealed highly conserved regions (Query 24–264 / Subject 21–255) with minimal gaps and several conservative substitutions, indicating structural and functional preservation. Based on this strong homology and the conservation of functional residues, it is likely that A0A8S4BTT0 shares a similar role in MHC class II-mediated antigen processing and presentation, suggesting a previously uncharacterized immunological function in *Menidia menidia*. Although CD74 is well characterized in mammals, the strong sequence conservation observed suggests that the Menidia menidia protein may perform a homologous immune-related role adapted to teleost fish biology.

[All Python scripts used for sequence analysis, BLAST execution, and homology-based functional annotation in this study have been uploaded to a public GitHub repository to ensure transparency and reproducibility.]