“*In silico* identification and functional characterisation of Type-Specific Antigen of 47kDa protein from *Orientia tsutsugamushi* using sequence analysis and homology-based annotation”

05.02.2026

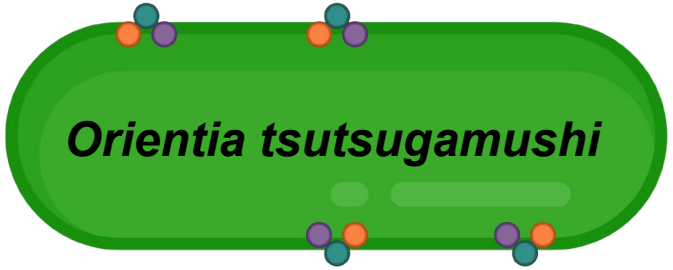
A Biopython-based pipeline collaborative project by:
Gayathri Snigdha, Paripoorna Reddy, Purvi Shah, Sameen Khalid, Siddhesh Uday Sapre, Vaidehi Kadam

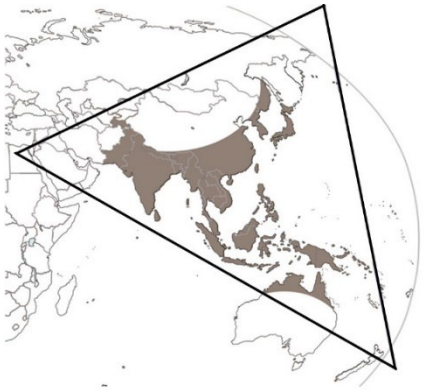
Guided by: Muskan Kashyap

Introduction-Disease and the Pathogen

Orientia tsutsugamushi (OT)

- Gram-negative obligate intracellular bacterium
- Structural rigidity conveyed by cross-linked outer membrane proteins, e.g. Type-specific antigen of 22 kD (TSA22), of 47 kD (TSA47) and 56kD (TSA56)[1,2]





Tsutsugamushi triangle

Scrub Typhus

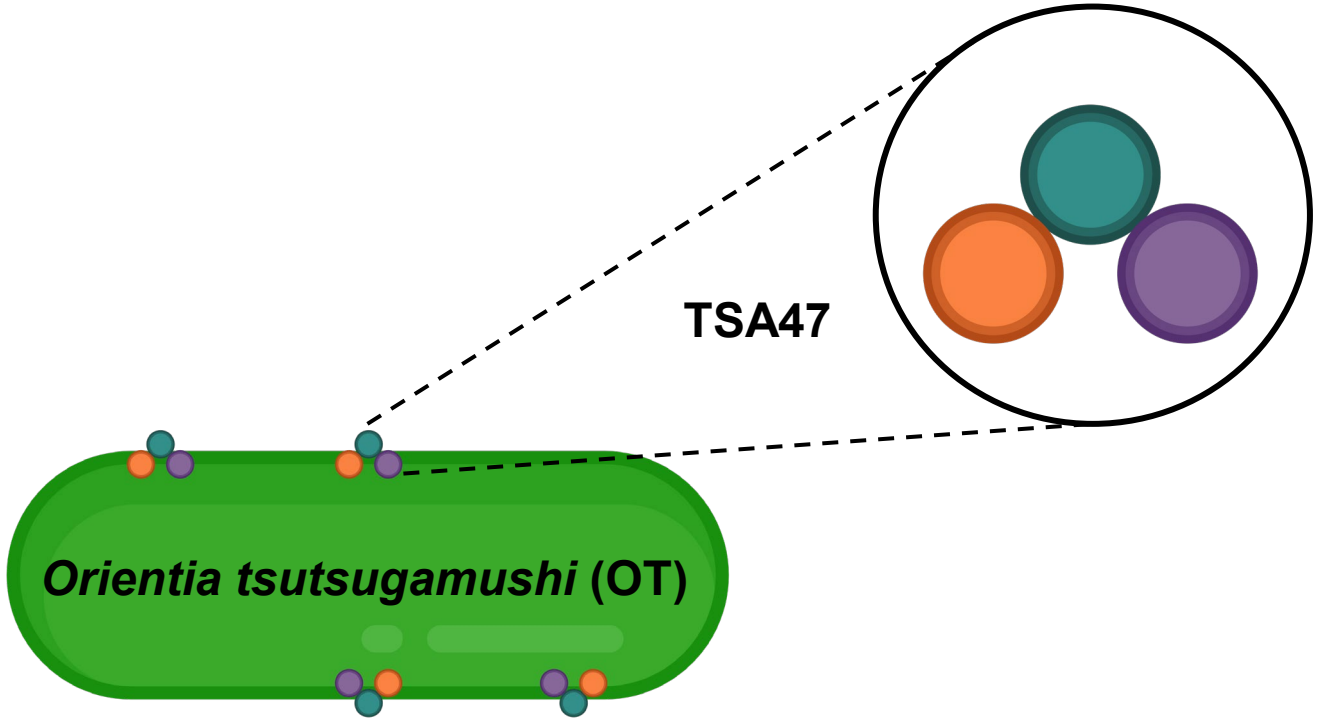
- Emerging neglected tropical disease : ~1 million cases/ year
- Endemicity: Asia-Pacific region
- Transmission: larvae of harvest mites (‘chiggers’)
- No vaccine available
- Treatment with antibiotics, e.g. Doxycycline

Mite (Vector)

[1]Lin CC *et al.*, 2012, *Int J Mol Med.*

[2]Fromm L. *et al.*, 2023, *Microbiol. Open*

Introduction-TSA47



- Highly conserved between strains (96.4-100%), involved in budding out of host cells[1]
- Understudied pertaining its role in OT life cycle

Objective: *In silico* dissection and extrapolation of the functional role of TSA47 using Biopython

[1]Kim *et al.*, 2013, *Microb. Pathog.*

Step 1: Sequence Retrieval

- NCBI Protein ID: SPR14422.1 retrieved from Nucleotide ID: LS398548

Input Syntax

Step1_FASTA_Format.py > ...

```python
#For this Project, in Step 1 (Biological Sequence Selection/Retrieval),we chose disease related protein.
#A gene from bacteria 'Orientia tsutsugamushi' (OT) which causes human disease 'Scrub typhus' was chosen.
#This gene encodes for a protein of the size 47kDa (molecular weight). It is known as TSA47 or bacterial htrA1.
#Biological role of TSA47 is not yet understood in OT's lif-cycle.
#Abbreviations: TSA: Type-Specific Antigen, htrA: High-Temperature Requirement Protein A
print("Step 1: performed by Siddhesh Uday Sapre")
print("-"*90)
from Bio import SeqIO
record = SeqIO.read("Otsutsugamushi_Karp_tsa47.fasta", "fasta")
print("The following protein record has Nucleotide ID:", record.id)
print(record.description)
print("The length of the sequence is:", len(record.seq), "amino acids")
print("The protein sequence is:", record.seq)
```

Results(Output)

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
"c:/Users/Siddhesh/Desktop/Biopython Pipeline Project-Group 2/Step1_FASTA_Format.py"
Step 1: performed by Siddhesh Uday Sapre
------------------------------------------------------------------------------------------
The following protein record has Nucleotide ID: SPR14422.1
SPR14422.1 Type surface antigen 47 (47 kDa/htrA) [Orientia tsutsugamushi]
The length of the sequence is: 466 amino acids
The protein sequence is: MKKAFYLHLIVFALQGISNVHSKSLLNQKALLPQQKSDMHINVNSLSDIVEPLISTVVSIYAVDTNIGISFNNKVSKYQQEVFLGSGVIIDSSGYIVTNENVIAGAENIKV
KLHDGSELIAELVGSDNKINIALLKINSPAALSYATFGDSNQSRVGDQVIAIGSPFGLRGTVTNGIISSKGRDMGNGIVTDFIQTNAAIHMGSFGGPMFNLEGKIIGINSIHVSYSGISFAIPSNTVLEAVECLKK
GEKIRRGMLNVMLNELTPELNENLGLKQDQNGVLITEVIKEGSAAQCGIAPGDVITKFHDKAIKTGRDLQVAVSSTMLNSEREVELLRNGKSMTLKCKIIANKGEDSEQQSNDQSLVVNGVKFVDLTPDLVKKYNI
TSANNNGLFVLEVSPNSSWGRYGLKMGLRPRDIILSVKRDDNKKDISVKTLREIVTNIKHNEIFFTVQRGDRMLYIALPNINK
PS C:\Users\Siddhesh\Desktop\Biopython Pipeline Project-Group 2>
```

The protein from OT (*aka* htrA) has a length of 466 amino acids

Step 2: Sequence Quality Analysis

Input Syntax- Part 1/2

Step2_FASTA_format.py > ...

```python
#Step 2 of the project has been performed to perform the Sequence Quality Analysis.
#This includes reading the sequence using BioPython, calculation of each amino acid content.
#In addition, I have also included the amino acid sequence percentage for different types of amino acid content.
from Bio import SeqIO
record = SeqIO.read("Otsutsugamushi_Karp_tsa47.fasta", "fasta")
Alanine_count_A = record.seq.count("A")
Cysteine_count_C = record.seq.count("C")
Aspartate_count_D = record.seq.count("D")
Glutamate_count_E = record.seq.count("E")
Phenylalanine_count_F = record.seq.count("F")
Glycine_count_G = record.seq.count("G")
Histidine_count_H = record.seq.count("H")
Isoleucine_count_I = record.seq.count("I")
Lysine_count_K = record.seq.count("K")
Leucine_count_L = record.seq.count("L")
Methionine_count_M = record.seq.count("M")
Aspargine_count_N = record.seq.count("N")
Proline_count_P = record.seq.count("P")
Glutamine_count_Q = record.seq.count("Q")
Arginine_count_R = record.seq.count("R")
Serine_count_S = record.seq.count("S")
Threonine_count_T = record.seq.count("T")
Valine_count_V = record.seq.count("V")
Tryptophan_count_W = record.seq.count("W")
Tyrosine_count_Y = record.seq.count("Y")
StopCodon_count =record.seq.count("*")
print("The query protein is", len(record.seq), "amino acids long.")
print("-"*90)
```

Step 2: Sequence Quality Analysis

Input Syntax- Part 2/2

Step2_FASTA_format.py > ...

```
29  print("Alanine count is:", Alanine_count_A)
30  print("Cystein count is:", Cysteine_count_C)
31  print("Aspartate count is:", Aspartate_count_D)
32  print("Glutamate count is:", Glutamate_count_E)
33  print("Phenylalanine count is.", Phenylalanine_count_F)
34  print("Glycine count is:", Glycine_count_G)
35  print("Histidine count is:", Histidine_count_H)
36  print("Isoleucine count is:", Isoleucine_count_I)
37  print("Lysine count is:", Lysine_count_K)
38  print("Leucine count is:", Leucine_count_L)
39  print("Mthionine count is:", Methionine_count_M)
40  print("Aspargine count is:", Aspargine_count_N)
41  print("Proline count is:", Proline_count_P)
42  print("Glutamine count is:", Glutamine_count_Q)
43  print("Arginine count is:", Arginine_count_R)
44  print("Serine count is:", Serine_count_S)
45  print("Threonine count is:", Threonine_count_T)
46  print("Valine count is:", Valine_count_V)
47  print("Tryptophan count is:", Tryptophan_count_W)
48  print("Tyrosine count is:", Tyrosine_count_Y)
49  print("% Non-polar Aliphatic/ Hydrophobic amino acids:", (Glycine_count_G + Alanine_count_A + Valine_count_V + Leucine_count_L + Isoleucir
50  print("% Non-polar Aromatic/ Hydrophobic amino acids:", (Phenylalanine_count_F + Tryptophan_count_W)/len(record.seq)*100)
51  print("% Polar Uncharged amino acids/ Hydrophilic:", (Serine_count_S + Threonine_count_T + Cysteine_count_C + Aspargine_count_N + Glutamir
52  print("% Acidic Negatively charged/ Hydrophilic amino acids:", (Aspartate_count_D + Glutamate_count_E)/len(record.seq)*100)
53  print("% Basic Positively charged/Hydrophilic amino acids:", (Lysine_count_K + Arginine_count_R + Histidine_count_H)/len(record.seq)*100)
54  print("-"*90)
55  print("Stop codons in this sequence (expected to be one):",StopCodon_count+1)
56  print("-"*90)
57  print("End of Quality check. All parameters fulfilled. The query sequence passed the quality checks successfully.")
```

Note: Complete syntax could not be included in the snapshot due to limitations of display in Micrososft Visual Studio Code

Step 2: Sequence Quality Analysis

Results(Output)

```
iddhesh/Desktop/Biopython Pipeline Project-Group 2/Step1_FASTA_Format.py"
Step 1: performed by Siddhesh Uday Sapre
--------------------------------------------------------------------------------
The following protein record has Nucleotide ID: SPR14422.1
SPR14422.1 Type surface antigen 47 (47 kDa/htrA) [Orientia tsutsugamushi]
The length of the sequence is: 466 amino acids
The protein sequence is: MKKAFYLHLIVFALQGISNVHSKSLLNQKALLPQQKSDMHINVNSLSDIVEPLISTVVSIYAVDTNIGISFNNKVSKYQQEVFLGSGVIIDSSGYIVTNENVIAGAENIKVKLHDGSELIAE
LVGSDNKINIALLKINSPAALSYATFGDSNQSRVGDQVIAIGSPFGLRGTVTNGIISSKGRDMGNGIVTDFIQTNAAIHMGSFGGPMFNLEGKIIGINSIHVSYSGISFAIPSNTVLEAVECLKKGEKIRRGMLNVMLNELTPELNE
NLGLKQDQNGVLITEVIKEGSAAQCGIAPGDVITKFHDKAIKTGRDLQVAVSSTMLNSEREVELLRNGKSMTLKCKIIANKGEDSEQQSNDQSLVVNGVKFVDLTPDLVKKYNITSANNNGLFVLEVSPNSSWGRYGLKMGLRPRDI
ILSVKRDDNKKDISVKTLREIVTNIKHNEIFFTVQRGDRMLYIALPNINK
PS C:\Users\Siddhesh\Desktop\Biopython Pipeline Project-Group 2> & C:/Users/Siddhesh/AppData/Local/Programs/Python/Python314/python.exe "c:/Users/S
iddhesh/Desktop/Biopython Pipeline Project-Group 2/Step2_FASTA_format.py"
The query protein is 466 amino acids long.
--------------------------------------------------------------------------------
Alanine count is: 24
Cystein count is: 3
Aspartate count is: 23
Glutamate count is: 23
Phenylalanine count is. 15
Glycine count is: 39
Histidine count is: 8
Isoleucine count is: 47
Lysine count is: 35
Leucine count is: 43
Mthionine count is: 11
Aspargine count is: 40
Proline count is: 12
Glutamine count is: 17
Arginine count is: 15
Serine count is: 42
Threonine count is: 20
Valine count is: 39
Tryptophan count is: 1
```

Step 2: Sequence Quality Analysis

Results(Output)

```
% Non-polar Aliphatic/ Hydrophobic amino acids: 46.137339055793994
% Non-polar Aromatic/ Hydrophobic amino acids: 3.4334763948497855
% Polar Uncharged amino acids/ Hydrophilic: 28.11158798283262
% Acidic Negatively charged/ Hydrophilic amino acids: 9.871244635193133
% Basic Positively charged/Hydrophilic amino acids: 12.446351931330472
--------------------------------------------------------------------------------
Stop codons in this sequence (expected to be one): 1
--------------------------------------------------------------------------------
End of Quality check. All parameters fulfilled. The query sequence passed the quality checks successfully.
PS C:\Users\Siddhesh\Desktop\Biopython Pipeline Project-Group 2>
```

1. All individual amino acids were enumerated within the protein

2. Percentages of different types of amino acids were enumerated within the protein

Step 2: Sequence Quality Analysis

To check the quality of the sequence:

1. Sequence length
2. Low quality bases
3. N-terminal and C-terminal residues
4. Internal stop codon

```python
from Bio import SeqIO

valid_aa = set("ACDEFGHIKLMNPQRSTVWY")

record = SeqIO.read("protein_endoprotease.fasta", "fasta")
sequence = str(record.seq)

low_quality_count = sum(1 for aa in sequence if aa not in valid_aa)


print("Protein ID:", record.id)
print("Length:", len(sequence))
print("Low-quality residues:", low_quality_count)

if "*" in sequence[:-1]:
    print("❌ Internal stop codon")
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS                                Python

PS C:\Users\HOME\Desktop\biopython course> & C:/Users/HOME/AppData/Local/Programs/Python/Python311/python.exe "c:/Users/HOME/Desktop/biopytho
n course/project_1"
Protein ID: tr|A0A2U3QMG0|A0A2U3QMG0_ORITS
Length: 432
Low-quality residues: 0
```

Sequence passes criteria 1,2,3 & 4→Query sequence was used further for analysis

STEP:3 SEQUENCE FILTERING AND VALIDITY

DONE BY: Purvi Shah

Presented by: Sameen Khalid

Step 3: Sequence Filtering & Validation

Input file: *Orientia tsutsugamushi* – TSA47 gene (1401 bp CDS)

Quality Control Workflow

DNA-level filtering

- Minimum length ≥ 90 bp
- Valid nucleotides only (A, T, G, C)
- Length divisible by 3 (correct reading frame)
- Must start with start codon (ATG)

Protein-level filtering

- Valid amino acids only
- No internal stop codons
- Stop codon allowed only at end

Filtered FASTA file

- Only QC-passed proteins saved
- Example: TSA47 protein (~467 aa)

Step3_Seq_Filter_and_Validation.py > ...

```python
from Bio import SeqIO
from Bio.SeqRecord import SeqRecord

VALID_DNA = set("ATGCN")
VALID_AA = set("ACDEFGHIKLMNPQRSTVWY*")

input_gb = "Otsutsugamushi_Karp_tsa47.gb"
output_fasta = "filtered_proteins.fasta"

filtered_records = []

for record in SeqIO.parse(input_gb, "genbank"):

    for feature in record.features:

        if feature.type == "CDS":

            # -------- DNA extraction --------
            dna_seq = str(feature.extract(record.seq)).upper()

            # ---- DNA QC ----
            if len(dna_seq) < 90:
                continue

            if not set(dna_seq).issubset(VALID_DNA):
                print(f"Invalid nucleotides in {record.id}")
                continue

            if len(dna_seq) % 3 != 0:
                print(f"Frame issue in {record.id}")
                continue

            if not dna_seq.startswith("ATG"):
                print(f"No start codon in {record.id}")
                continue
```

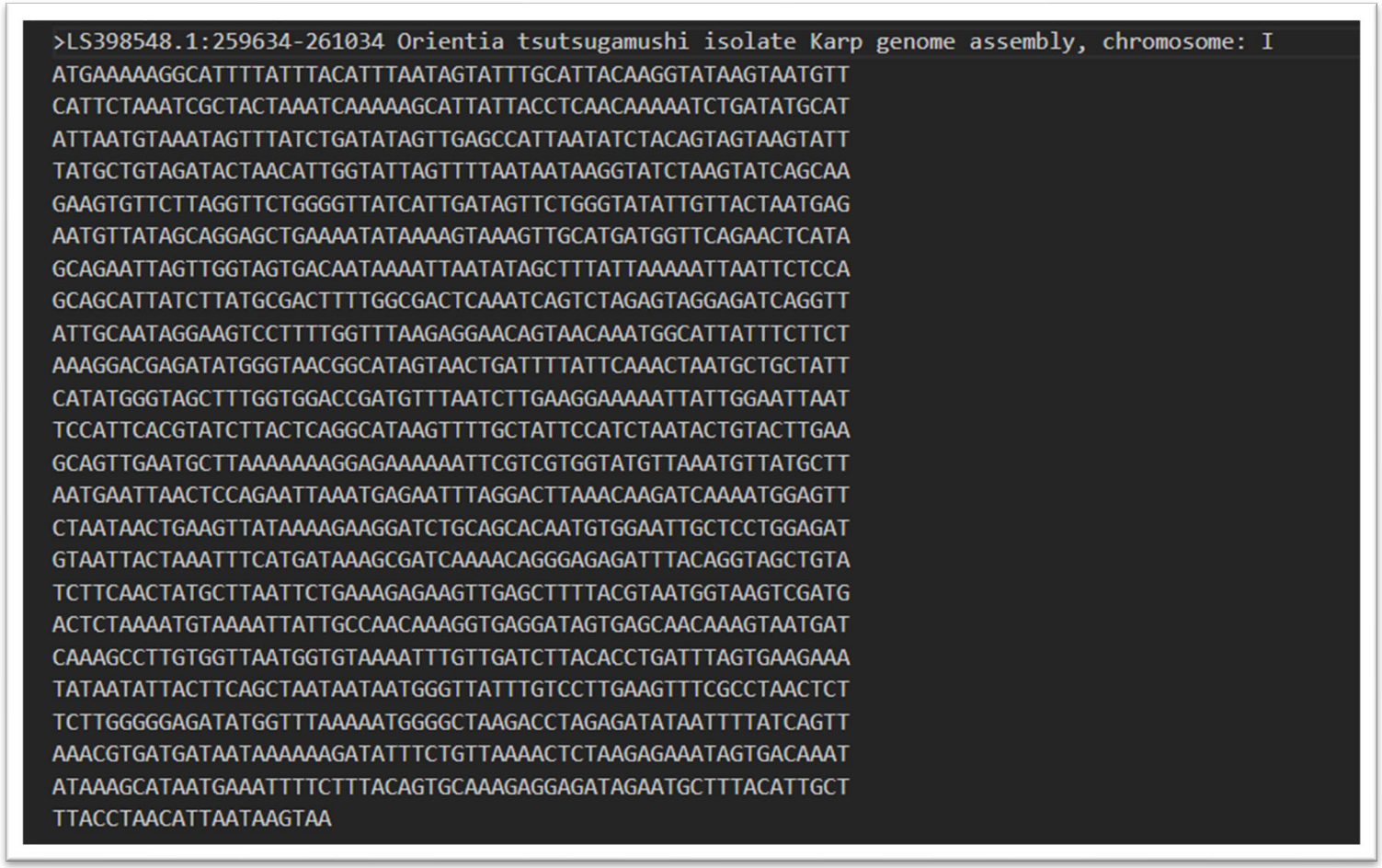
Step3_Seq_Filter_and_Validation.py > ...

```python
            # ---- Protein QC ----
            if not set(protein_seq).issubset(VALID_AA):
                print(f"Invalid AA in {record.id}")
                continue

            if "*" in protein_seq[:-1]:
                print(f"Internal stop codon in {record.id}")
                continue

            # Save
            protein_record = SeqRecord(
                feature.extract(record.seq).translate(table=11),
                id=record.id,
                description="CDS protein (DNA+Protein QC passed)"
            )

            filtered_records.append(protein_record)

    SeqIO.write(filtered_records, output_fasta, "fasta")

    print(f"Saved {len(filtered_records)} high-quality proteins.")

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

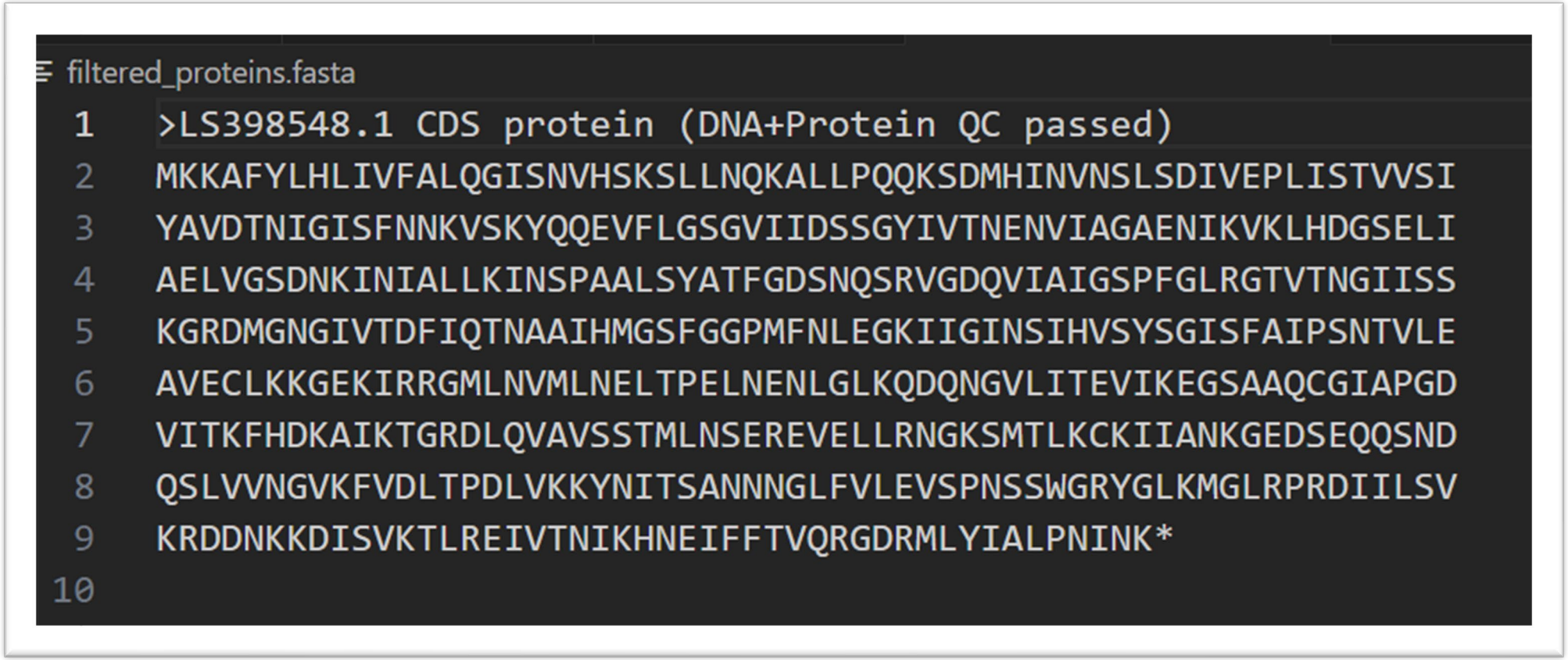
Python

```
PS C:\Users\purvi\Desktop\BioPython> & C:/Users/purvi/AppData/Local/Python/pythoncore-3.14-64/python.exe c:/Users/purvi/Desktop/BioPython/Step3_Seq_Filter_and_
alidation.py
Saved 1 high-quality proteins.
PS C:\Users\purvi\Desktop\BioPython>
```



```
>LS398548.1:259634-261034 Orientia tsutsugamushi isolate Karp genome assembly, chromosome: I
ATGAAAAAGGCATTTTATTTACATTTAATAGTATTTGCATTACAAGGTATAAGTAATGTT
CATTCTAAATCGCTACTAAATCAAAAAGCATTATTACCTCAACAAAAATCTGATATGCAT
ATTAATGTAAATAGTTTATCTGATATAGTTGAGCCATTAATATCTACAGTAGTAAGTATT
TATGCTGTAGATACTAACATTGGTATTAGTTTTAATAATAAGGTATCTAAGTATCAGCAA
GAAGTGTTCTTAGGTTCTGGGGTTATCATTGATAGTTCTGGGTATATTGTTACTAATGAG
AATGTTATAGCAGGAGCTGAAAATATAAAAGTAAAGTTGCATGATGGTTCAGAACTCATA
GCAGAATTAGTTGGTAGTGACAATAAAATTAATATAGCTTTATTAAAAATTAATTCTCCA
GCAGCATTATCTTATGCGACTTTTGGCGACTCAAATCAGTCTAGAGTAGGAGATCAGGTT
ATTGCAATAGGAAGTCCTTTTGGTTTAAGAGGAACAGTAACAAATGGCATTATTTCTTCT
AAAGGACGAGATATGGGTAACGGCATAGTAACTGATTTTATTCAAACTAATGCTGCTATT
CATATGGGTAGCTTTGGTGGACCGATGTTTAATCTTGAAGGAAAAATTATTGGAATTAAT
TCCATTCACGTATCTTACTCAGGCATAAGTTTTGCTATTCCATCTAATACTGTACTTGAA
GCAGTTGAATGCTTAAAAAAAGGAGAAAAAATTCGTCGTGGTATGTTAAATGTTATGCTT
AATGAATTAACTCCAGAATTAAATGAGAATTTAGGACTTAAACAAGATCAAAATGGAGTT
CTAATAACTGAAGTTATAAAAGAAGGATCTGCAGCACAATGTGGAATTGCTCCTGGAGAT
GTAATTACTAAATTTCATGATAAAGCGATCAAAACAGGGAGAGATTTACAGGTAGCTGTA
TCTTCAACTATGCTTAATTCTGAAAGAGAAGTTGAGCTTTTACGTAATGGTAAGTCGATG
ACTCTAAAATGTAAAATTATTGCCAACAAAGGTGAGGATAGTGAGCAACAAAGTAATGAT
CAAAGCCTTGTGGTTAATGGTGTAAAATTTGTTGATCTTACACCTGATTTAGTGAAGAAA
TATAATATTACTTCAGCTAATAATAATGGGTTATTTGTCCTTGAAGTTTCGCCTAACTCT
TCTTGGGGGAGATATGGTTTAAAAATGGGGCTAAGACCTAGAGATATAATTTTATCAGTT
AAACGTGATGATAATAAAAAAGATATTTCTGTTAAAACTCTAAGAGAAATAGTGACAAAT
ATAAAGCATAATGAAATTTTCTTTACAGTGCAAAGAGGAGATAGAATGCTTTACATTGCT
TTACCTAACATTAATAAGTAA
```

Filtered DNA sequence



Filtered Protein Sequence

STEP:4 HOMOLOGY SEARCH (BLAST)

Done BY: Sameen Khalid

Presented by: Sameen Khalid

Step 4: Homology Search (BLAST)

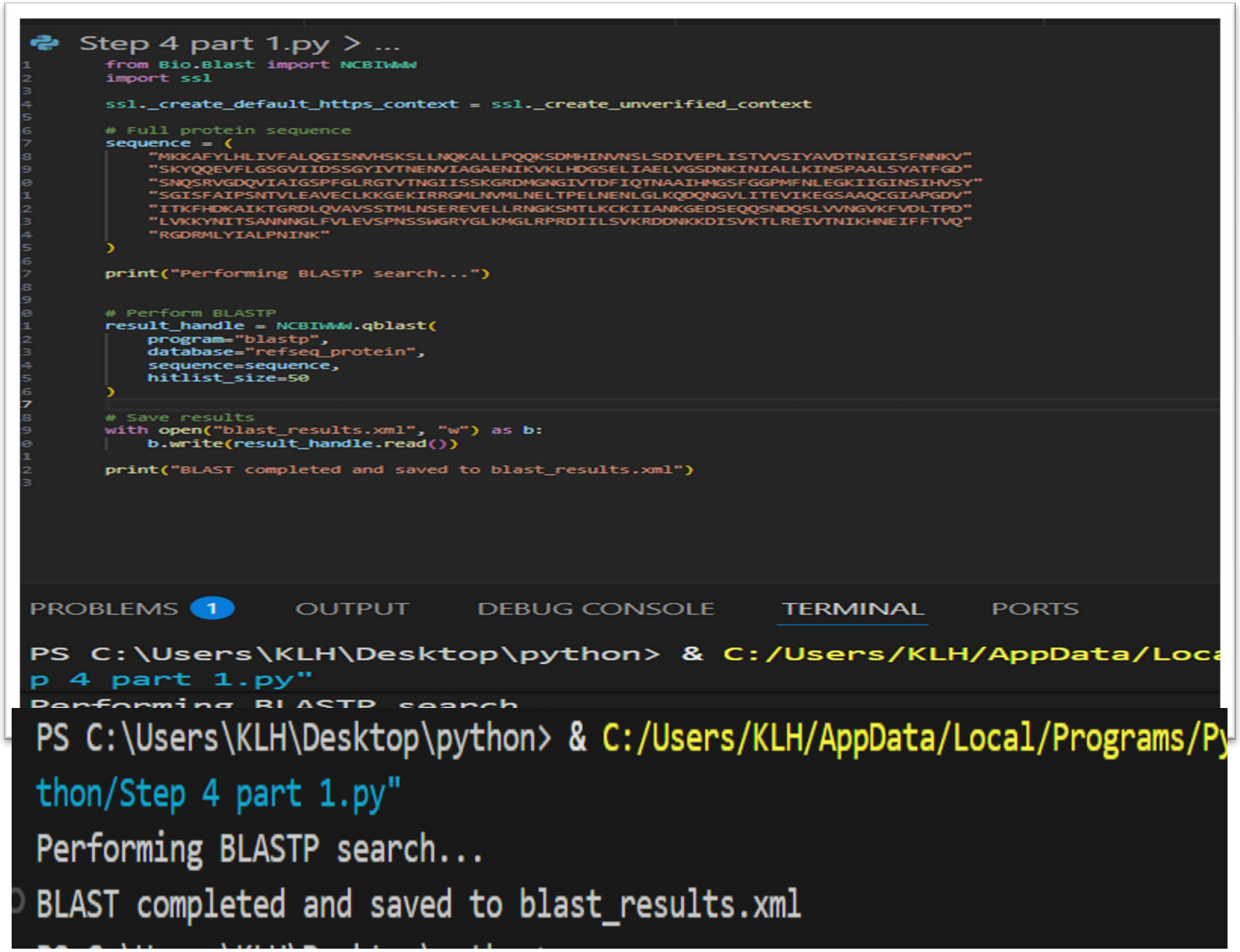
Identify:

1)Similarity search using BLAST:

2) Closest homologs, Conserved regions, Evolutionary hints

Similarity Search

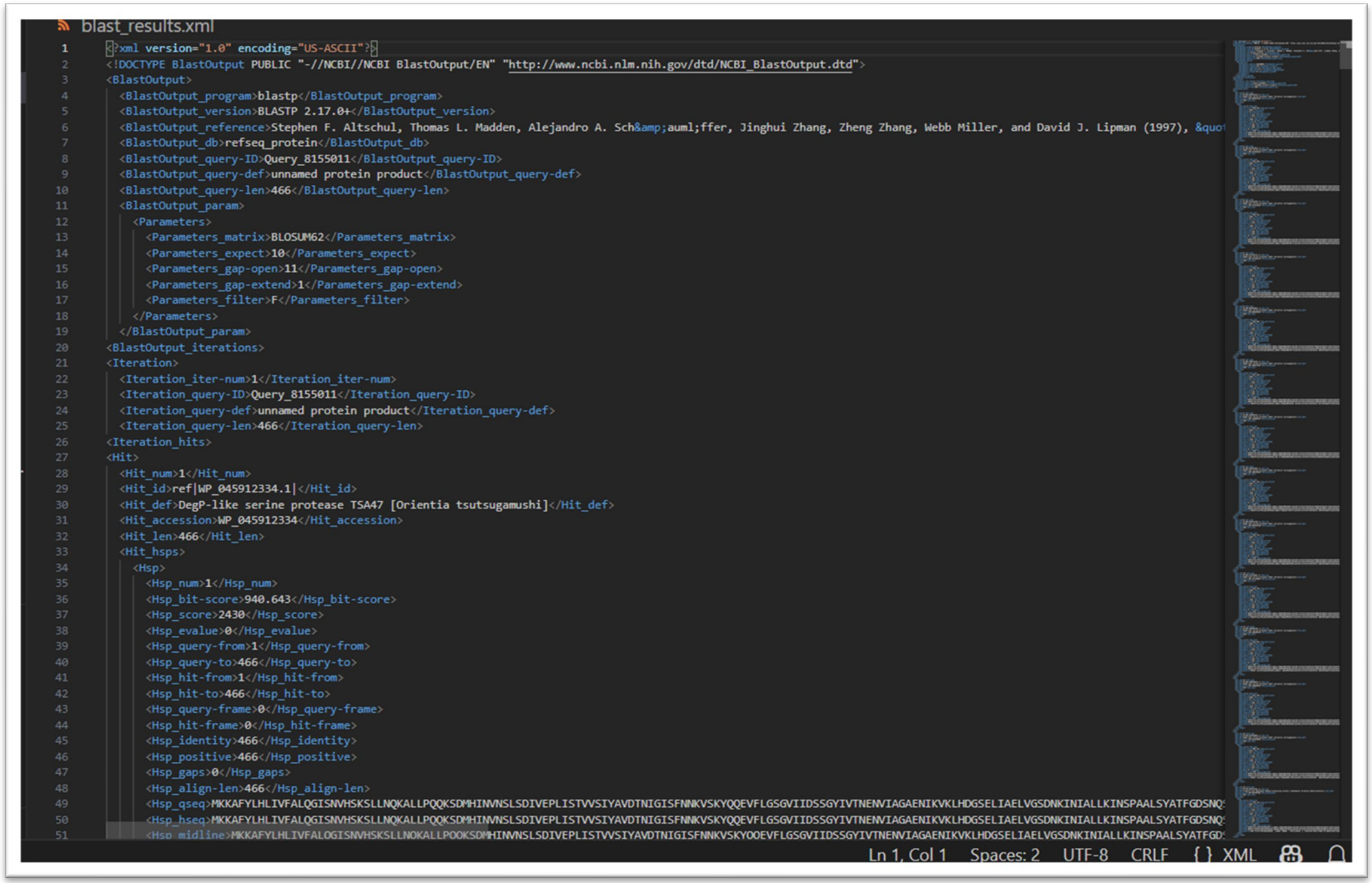
- Using the filtered protein sequence after sequence filtration and validity,
- A BLASTP similarity search was performed against the RefSeq protein database.
- Blast performed successfully and results were saved as blast_results.xml



Input and Output Syntax

Blast performed successfully

blast_results.xml



```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE BlastOutput PUBLIC "-//NCBI//NCBI BlastOutput/EN" "http://www.ncbi.nlm.nih.gov/dtd/NCBI_BlastOutput.dtd">
<BlastOutput>
  <BlastOutput_program>blastp</BlastOutput_program>
  <BlastOutput_version>BLASTP 2.17.0+</BlastOutput_version>
  <BlastOutput_reference>Stephen F. Altschul, Thomas L. Madden, Alejandro A. Sch&amp;auml;ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), &quot
  <BlastOutput_db>refseq_protein</BlastOutput_db>
  <BlastOutput_query-ID>Query_8155011</BlastOutput_query-ID>
  <BlastOutput_query-def>unnamed protein product</BlastOutput_query-def>
  <BlastOutput_query-len>466</BlastOutput_query-len>
  <BlastOutput_param>
    <Parameters>
      <Parameters_matrix>BLOSUM62</Parameters_matrix>
      <Parameters_expect>10</Parameters_expect>
      <Parameters_gap-open>11</Parameters_gap-open>
      <Parameters_gap-extend>1</Parameters_gap-extend>
      <Parameters_filter>F</Parameters_filter>
    </Parameters>
  </BlastOutput_param>
<BlastOutput_iterations>
<Iteration>
  <Iteration_iter-num>1</Iteration_iter-num>
  <Iteration_query-ID>Query_8155011</Iteration_query-ID>
  <Iteration_query-def>unnamed protein product</Iteration_query-def>
  <Iteration_query-len>466</Iteration_query-len>
<Iteration_hits>
<Hit>
  <Hit_num>1</Hit_num>
  <Hit_id>ref|WP_045912334.1|</Hit_id>
  <Hit_def>DegP-like serine protease TSA47 [Orientia tsutsugamushi]</Hit_def>
  <Hit_accession>WP_045912334</Hit_accession>
  <Hit_len>466</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>940.643</Hsp_bit-score>
      <Hsp_score>2430</Hsp_score>
      <Hsp_evalue>0</Hsp_evalue>
      <Hsp_query-from>1</Hsp_query-from>
      <Hsp_query-to>466</Hsp_query-to>
      <Hsp_hit-from>1</Hsp_hit-from>
      <Hsp_hit-to>466</Hsp_hit-to>
      <Hsp_query-frame>0</Hsp_query-frame>
      <Hsp_hit-frame>0</Hsp_hit-frame>
      <Hsp_identity>466</Hsp_identity>
      <Hsp_positive>466</Hsp_positive>
      <Hsp_gaps>0</Hsp_gaps>
      <Hsp_align-len>466</Hsp_align-len>
      <Hsp_qseq>MKKAFYLHLIVFALQGISNVHSKSLLNQKALLPQQKSDMHINVNSLSDIVEPLISTVVSIYAVDTNIGISFNNKVSKYQQEVFLGSGVIIDSSGYIVTNENVIAGAENIKVKLHDGSELIAELVGSDNKINIALLKINSPAALSYATFGDSNQ
      <Hsp_hseq>MKKAFYLHLIVFALQGISNVHSKSLLNQKALLPQQKSDMHINVNSLSDIVEPLISTVVSIYAVDTNIGISFNNKVSKYQQEVFLGSGVIIDSSGYIVTNENVIAGAENIKVKLHDGSELIAELVGSDNKINIALLKINSPAALSYATFGDSNQ
      <Hsp_midline>MKKAFYLHLIVFALQGISNVHSKSLLNQKALLPQQKSDMHINVNSLSDIVEPLISTVVSIYAVDTNIGISFNNKVSKYQQEVFLGSGVIIDSSGYIVTNENVIAGAENIKVKLHDGSELIAELVGSDNKINIALLKINSPAALSYATFGDS
```

Ln 1, Col 1 Spaces: 2 UTF-8 CRLF { } XML

Blast_results.xml

Homology Search & Sequence Similarity Analysis

Objective

- To identify similar proteins of the from *Orientia tsutsugamushi*

Method

- BLAST results were parsed using **Bio-python**
- Parameters were:

1. Homolog
2. Length
3. E-value
4. HSP score
5. Identity
6. Query region
7. Subject region

```python
from Bio.Blast import NCBIXML

import ssl

#from Bio import Entrez
#Entrez.email = "sameenkhaled@gmail.com"

ssl._create_default_https_context = ssl._create_unverified_context

# Step 1: Parse BLAST results
with open("blast_results.xml") as b:
    blast_record= NCBIXML.read(b)


print(len(blast_record.alignments))
print("The total alignments are:", "50")
#for alignment in blast_record.alignments:

for alignment in blast_record.alignments:
    print("\nHomolog:", alignment.title)
    print("Length:", alignment.length)

    for hsp in alignment.hsps:
        print(" E-value:", hsp.expect)
        print(" HSP score:", hsp.bits)
        print(" Identity:", hsp.identities, "/", hsp.align_length)
        print(" Query region:", hsp.query_start, "-", hsp.query_end)
        print(" Subject region:", hsp.sbjct_start, "-", hsp.sbjct_end)
```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\KLH\Desktop\python> & C:/Users/KLH/AppData/Local/Programs/Python/Python314/
y
50
The total alignments are: 50

Homolog: ref|WP_045912334.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
Length: 466
 E-value: 0.0
 HSP score: 940.643
 Identity: 466 / 466
 Query region: 1 - 466
 Subject region: 1 - 466

Homolog: ref|WP_012461825.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
```

Ln 31, Col 1

Homologs

Identification of Closest Homologs (Results)

50 closest homologs were identified

- Top hits showed:
  - **98–100% sequence identity**
  - **E-value = 0.0**
  - Full-length alignment (Query region: 1–466 aa)

Interpretation

- Extremely low E-values and high identity indicate **strong homology**
- Confirms correct identification of the target protein

step 4 part 3 > ...

```python
from Bio.Blast import NCBIXML
from Bio import Entrez, SeqIO
import ssl
from Bio.SeqRecord import SeqRecord
ssl._create_default_https_context = ssl._create_unverified_context

from collections import Counter

Entrez.email = "sameenkhaled@gmail.com"

from Bio.Blast import NCBIXML
# Step 1: Parse the BLAST results
with open("blast_results.xml") as handle:
    blast_records = list(NCBIXML.parse(handle))

blast_record = blast_records[0]

# Step 2: Closest homologs based on E-value and identity
closest_hits = []

for alignment in blast_record.alignments:
    best_hsp = max(alignment.hsps, key=lambda h: h.bits)
    identity_fraction = best_hsp.identities / best_hsp.align_length

    # for significant hits
    if best_hsp.expect < 1e-5 and identity_fraction >= 0.35:
        closest_hits.append({
            "title": alignment.title,
            "accession": alignment.accession,
            "identity": identity_fraction,
            "evalue": best_hsp.expect,
            "q_start": best_hsp.query_start,
            "q_end": best_hsp.query_end
        })

print(f"Number of closest homologs found: {len(closest_hits)}\n")

print("Top homologs (similarity info):")
for hit in closest_hits[:10]:
    print(f"- {hit['title']}")
    print(f"  Identity: {hit['identity']:.2%}")
    print(f"  E-value: {hit['evalue']}")
    print(f"  Query region: {hit['q_start']}-{hit['q_end']}\n")
```

step 4 part 3 > ...

```python
# Step 3: top hits from NCBI
records = []
print("Fetching sequences of top hits from NCBI...")
for hit in closest_hits[:10]:  # top 10 hits
    try:
        handle = Entrez.efetch(
            db="protein",
            id=hit["accession"],
            rettype="fasta",
            retmode="text"
        )
        record = SeqIO.read(handle, "fasta")
        records.append(record)
    except Exception as e:
        print(f"Failed to fetch {hit['accession']}: {e}")


SeqIO.write(records, "homologs.fasta", "fasta")
print("Sequences saved to homologs.fasta\n")

# Step 4: Conserved regions
if len(records) > 1:
    seq_length = min(len(r.seq) for r in records)
    conserved_positions = []

    for i in range(seq_length):
        column = [str(r.seq[i]) for r in records]
        most_common = Counter(column).most_common(1)[0]
        conservation = most_common[1] / len(records)
        if conservation >= 0.8:  # >=80% conserved
            conserved_positions.append((i + 1, most_common[0]))

    print("Highly conserved positions (1-based index):")
    print(conserved_positions[:20], "...")
else:
    print("Not enough sequences for conserved region analysis.\n")

# Step 5: Evolutionary hints (organism info from top hits)
print("\nEvolutionary hints (species/organism info from top hits):")
for hit in closest_hits[:10]:
    print(f"- {hit['title']}")
```

Commands section for Identifying Closest homologs, conserved regions and Evolutionary hints

Closest homologs

```
Number of closest homologs found: 50

Top homologs (similarity info):
- ref|WP_045912334.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 100.00%
  E-value: 0.0
  Query region: 1-466

- ref|WP_012461825.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.71%
  E-value: 0.0
  Query region: 1-466

- ref|WP_371222284.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.71%
  E-value: 0.0
  Query region: 1-466

- ref|WP_371253172.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.93%
  E-value: 0.0
  Query region: 1-466

- ref|WP_064644053.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.71%
  E-value: 0.0
  Query region: 1-466

- ref|WP_109227263.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.50%
  E-value: 0.0
  Query region: 1-466
```

```
- ref|WP_146695787.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.50%
  E-value: 0.0
  Query region: 1-466

- ref|WP_109234584.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.50%
  E-value: 0.0
  Query region: 1-466

- ref|WP_045914760.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.71%
  E-value: 0.0
  Query region: 1-466

- ref|WP_371219507.1| DegP-like serine protease TSA47 [Orientia tsutsugamushi]
  Identity: 98.50%
  E-value: 0.0
  Query region: 1-466
```

Top 10 Closest Homologs

Functional Annotation from Homology

Annotation Results

- All top homologs were annotated as:
  - **DegP-like serine protease TSA47**
  - Species: *Orientia tsutsugamushi*

Inference

- TSA47 is a **highly conserved DegP-like serine protease**

Conserved Region Analysis

```
Highly conserved positions (1-based index):
[(1, 'M'), (2, 'K'), (3, 'K'), (4, 'A'), (5, 'F'), (6, 'Y'), (7, 'L'), (8, 'H'), (9, 'L'), (10, 'I'), (11, 'V'), (12, 'F'
, (13, 'A'), (14, 'L'), (15, 'Q'), (16, 'G'), (17, 'I'), (18, 'S'), (19, 'N'), (20, 'V')] ...
```

Highly conserved regions

Key Finding

- Several highly conserved amino acids detected

Biological Significance

- Conserved regions suggest:
  - Functional importance
  - Structural stability
  - Possible role in host–pathogen interaction

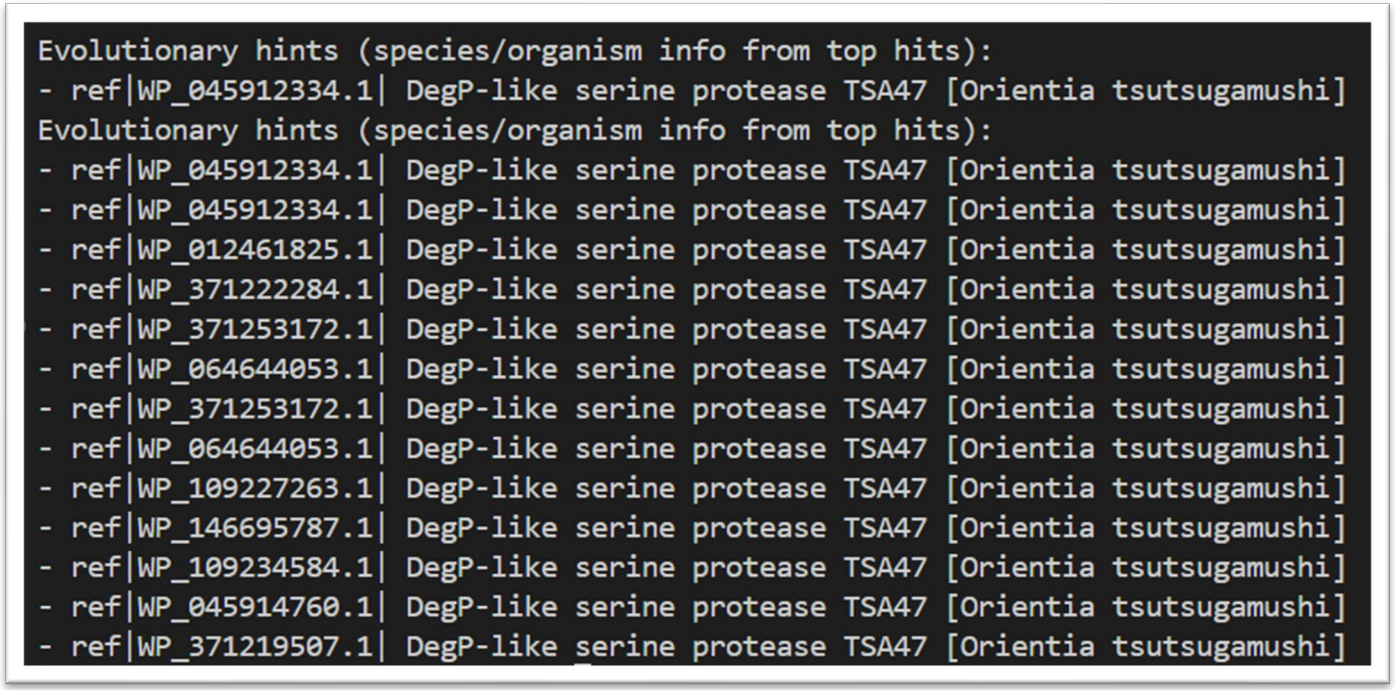
Evolutionary Insights

Observation

- All closest homologs belong to *Orientia tsutsugamushi*
- No distant species detected among top hits

Interpretation

- TSA47 is **species-specific**
- Shows **low evolutionary divergence**



Evolutionary hints

Step 4 insights

- A BLASTP similarity search was performed against the RefSeq protein database.
- The closest homologs were identified based on lowest E-values and highest percentage identity.
- Conserved regions were determined by comparing aligned homologous sequences.
- The taxonomic distribution of homologs provided evolutionary insights into the protein's conservation.

"**Homology-based analysis confirms that TSA47 is a highly conserved, functionally important protein in *Orientia tsutsugamushi*.**"

Step:5 Functional Annotation

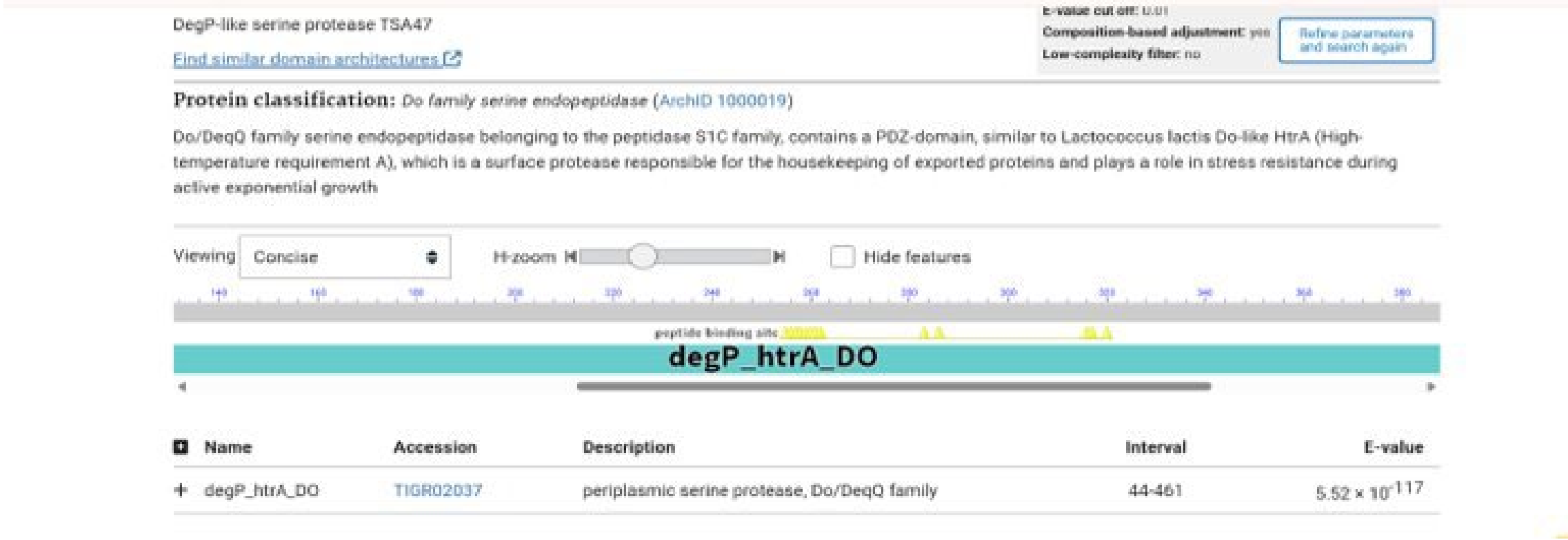
Done by: Gayathri Snighda

Presented by: Sameen Khalid

Step 5: Functional Annotation

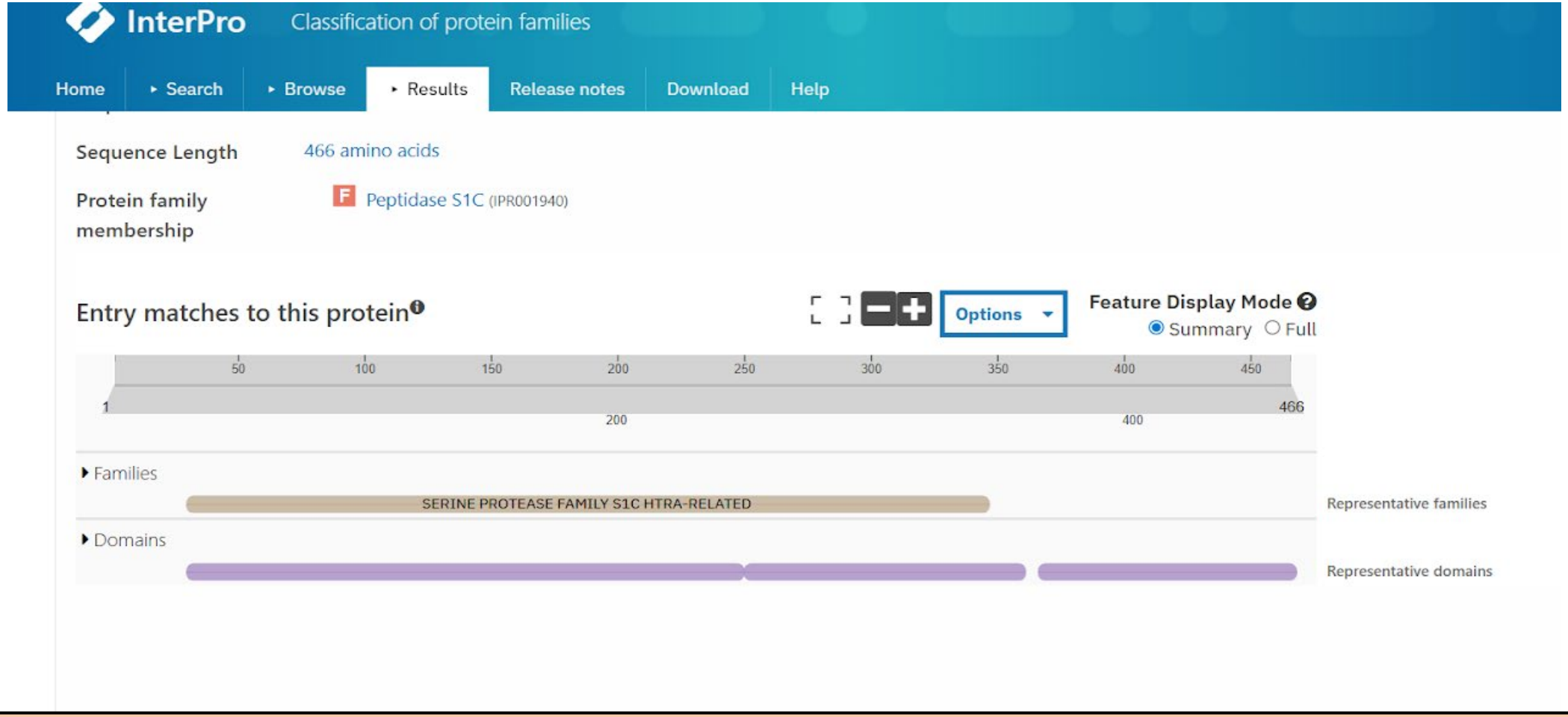
- Protein: Serine protease
- Function : a surface protease responsible for the housekeeping of exported proteins and plays a role in stress resistance during active exponential growth
- Organism: *Orientia tsutsugamushi* Karp

1.Conserved domains -NCBI CDD



Name	Accession	Description	Interval	E-value
degP_htrA_DO	TIGR02037	periplasmic serine protease, Do/DeqQ family	44-461	5.52×10^{-117}

Conserved Domain Database (CDD) analysis revealed the presence of a **degP_htrA_DO** spanning amino acids **44-461**, with a highly significant E-value (<1e-50), indicating a complete and functional protein.



InterPro analysis shows that the 466-aa protein belongs to the HtrA-related serine protease family (Peptidase S1C) and contains multiple domains: 2 PDZ and one trypsin like serine protease domain.

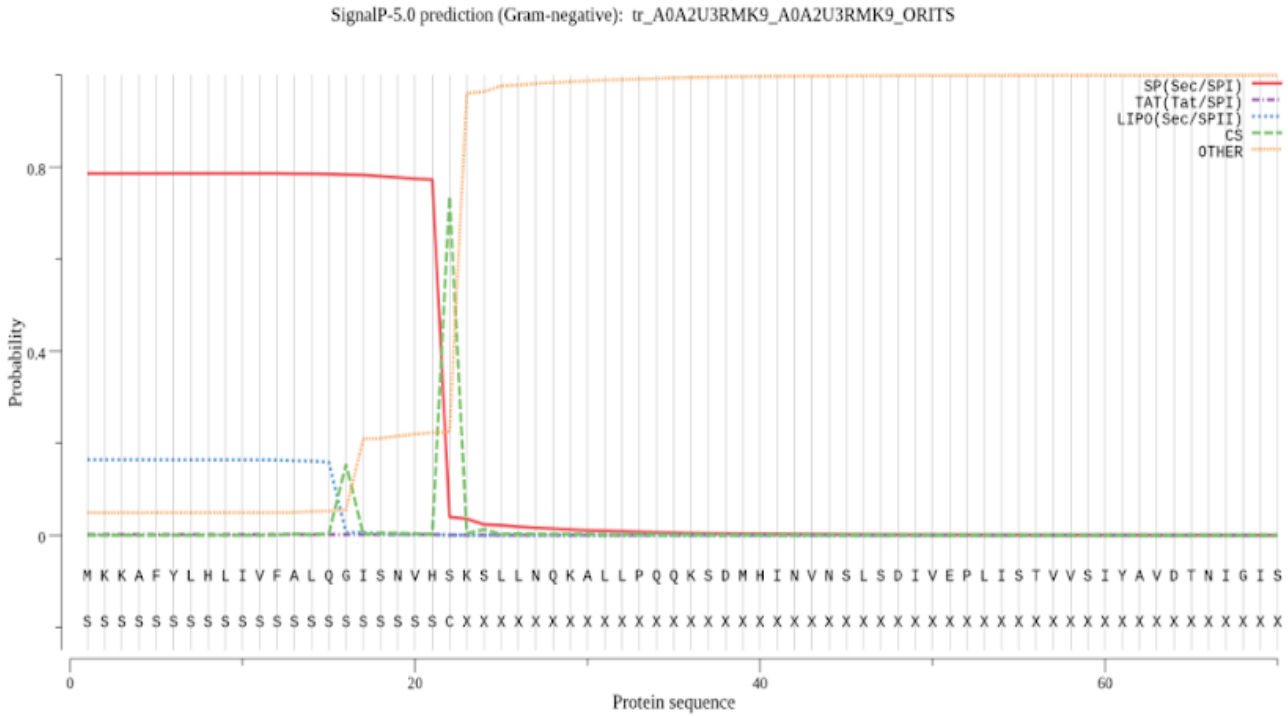
tr_A0A2U3RMK9_A0A2U3RMK9_ORITS

Prediction: Signal peptide (Sec/SPI)

Cleavage site between pos. 22 and 23: VHS-KS. Probability: 0.7358

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.7848	0.0015	0.1637	0.05

Download: PNG / EPS / Tabular

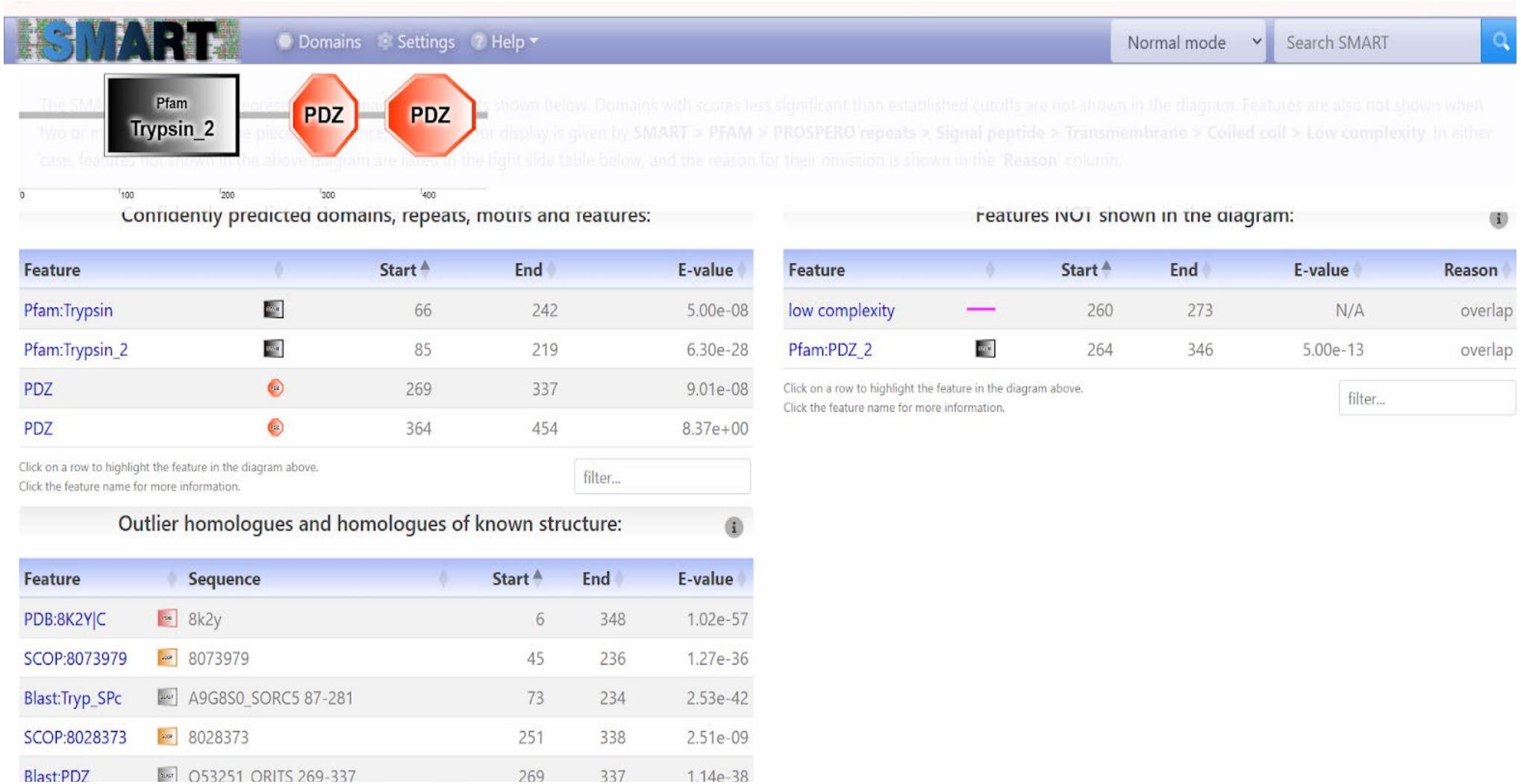


The protein is a classical Sec-dependent secreted protein. The first 22 amino acids act as a signal peptide, which are cleaved off. The rest of the protein is mature and functional. It is **not cytoplasmic** and not a lipoprotein.

Domain analysis

PSORTb Results (Click here for an explanation of the output formats)

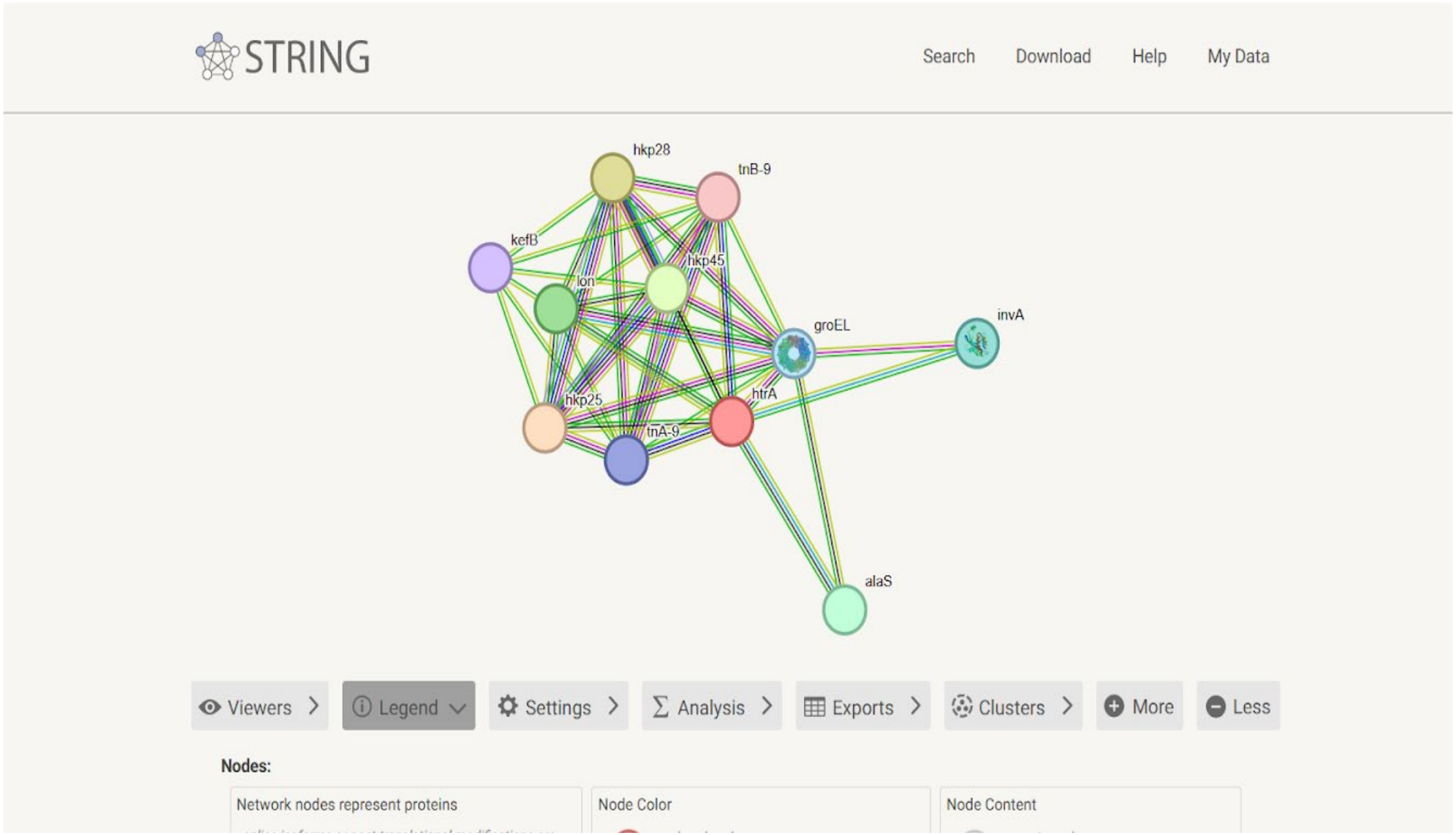
```
SeqID: tr|A0A2U3RMK9|A0A2U3RMK9_ORITS Probable periplasmic serine endoprotease DegP-like OS=Orientia tsutsugamushi OX=784 GN=TSA47 PE=3 SV=1
  Analysis Report:
    CMSVM-            Unknown                  [No details]
    CytoSVM-          Unknown                  [No details]
    ECSVM-            Unknown                  [No details]
    ModHMM-           Unknown                  [No internal helices found]
    Motif-            Unknown                  [No motifs found]
    OMPMotif-         Unknown                  [No motifs found]
    OMSVM-            Unknown                  [No details]
    PPSVM-            Unknown                  [No details]
    Profile-          Unknown                  [No matches to profiles found]
    SCL-BLAST-        Periplasmic              [matched 15595963: serine protease MucD precursor[Pseudomonas aeruginosa PAO1]]
    SCL-BLASTe-       Unknown                  [No matches against database]
    Signal-           Unknown                  [No signal peptide detected]
  Localization Scores:
    Cytoplasmic           0.33
    CytoplasmicMembrane   0.06
    Periplasmic           9.44
    OuterMembrane         0.06
    Extracellular         0.11
  Final Prediction:
    Periplasmic           9.44
----------------------------------------------------------------------------
```

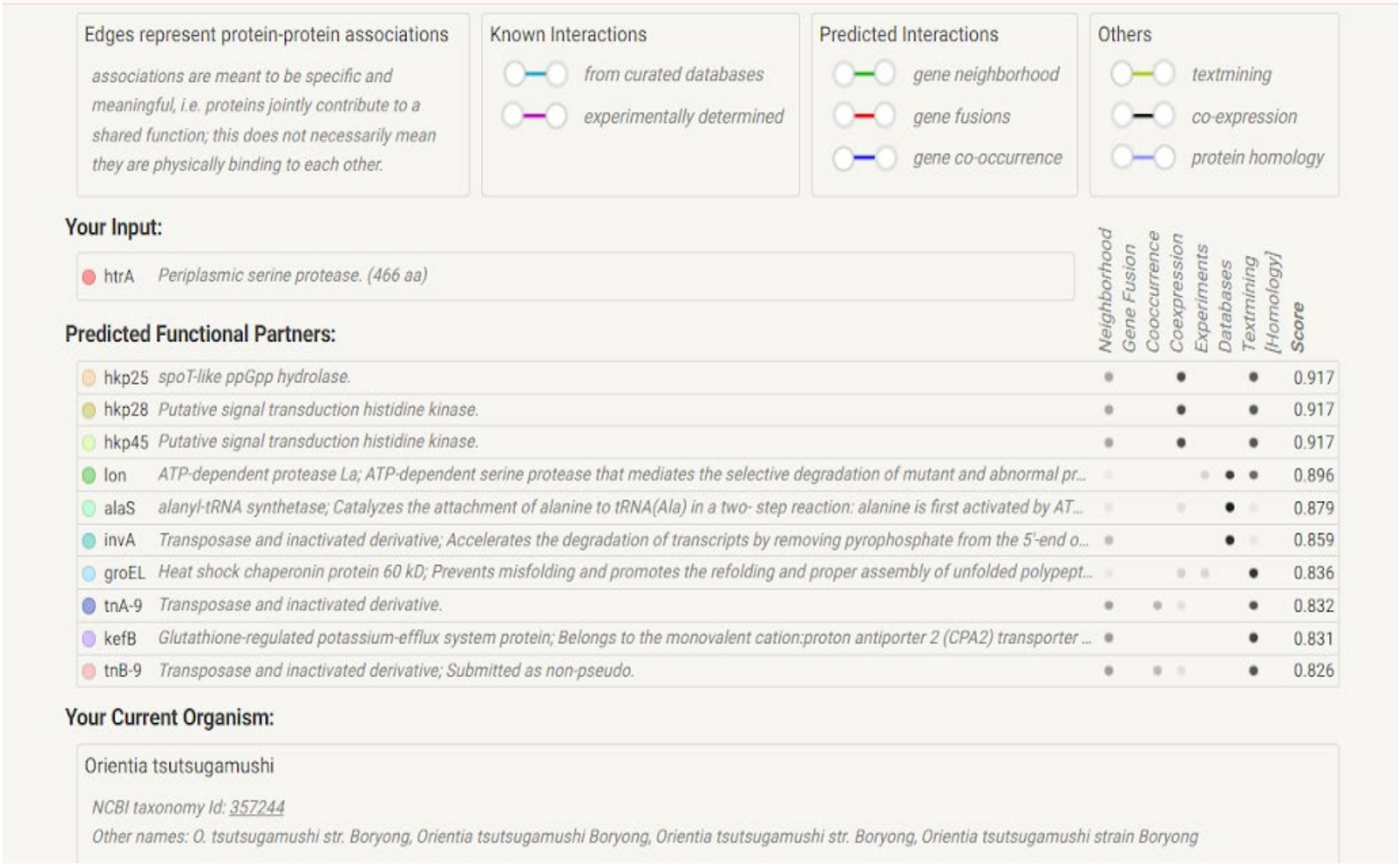

SMART
Domains
Settings
Help
Normal mode
Search SMART
Pfam
Trypsin_2
PDZ
PDZ
0
100
200
300
400
Confidently predicted domains, repeats, motifs and features:
Features NOT shown in the diagram:
Feature
Start
End
E-value
Pfam:Trypsin
66
242
5.00e-08
Pfam:Trypsin_2
85
219
6.30e-28
PDZ
269
337
9.01e-08
PDZ
364
454
8.37e+00
Click on a row to highlight the feature in the diagram above.
Click the feature name for more information.
filter...
Feature
Start
End
E-value
Reason
low complexity
260
273
N/A
overlap
Pfam:PDZ_2
264
346
5.00e-13
overlap
Click on a row to highlight the feature in the diagram above.
Click the feature name for more information.
filter...
Outlier homologues and homologues of known structure:
Feature
Sequence
Start
End
E-value
PDB:8K2Y|C
8k2y
6
348
1.02e-57
SCOP:8073979
8073979
45
236
1.27e-36
Blast:Tryp_SPc
A9G8S0_SORC5 87-281
73
234
2.53e-42
SCOP:8028373
8028373
251
338
2.51e-09
Blast:PDZ
O53251_ORITS 269-337
269
337
1.14e-38

Trypsin domain (aa 66–242) and **Trypsin_2 domain (aa 85–219)** indicate that the protein has a **serine protease function**,Two **PDZ domains** were detected: the first (aa 269–337) with high confidence (E-value 9.01e-08), and the second (aa 364–454) with lower confidence (E-value 8.37).

Collectively, the mature protein appears to contain a **functional protease domain** and **protein interaction modules**.

Homology searches identified several structural and sequence homologues of the protein. The Trypsin domain shows strong similarity to experimentally characterized serine proteases (e.g., PDB:8K2Y, SCOP:8073979)


STRING
Search
Download
Help
My Data
hkp28
tnB-9
kefB
hkp45
lon
groEL
invA
hkp25
htrA
tnA-9
alaS
Viewers
Legend
Settings
Analysis
Exports
Clusters
More
Less
Nodes:
Network nodes represent proteins
Node Color
Node Content



Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

- *from curated databases*
- *experimentally determined*

Predicted Interactions

- *gene neighborhood*
- *gene fusions*
- *gene co-occurrence*

Others

- *textmining*
- *co-expression*
- *protein homology*

Your Input:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
htrA	*Periplasmic serine protease. (466 aa)*									

Predicted Functional Partners:

hkp25	*spoT-like ppGpp hydrolase.*	●			●			●		0.917
hkp28	*Putative signal transduction histidine kinase.*	●			●			●		0.917
hkp45	*Putative signal transduction histidine kinase.*	●			●			●		0.917
lon	*ATP-dependent protease La; ATP-dependent serine protease that mediates the selective degradation of mutant and abnormal pr...*					●	●	●		0.896
alaS	*alanyl-tRNA synthetase; Catalyzes the attachment of alanine to tRNA(Ala) in a two- step reaction: alanine is first activated by AT...*						●			0.879
invA	*Transposase and inactivated derivative; Accelerates the degradation of transcripts by removing pyrophosphate from the 5'-end o...*	●					●			0.859
groEL	*Heat shock chaperonin protein 60 kD; Prevents misfolding and promotes the refolding and proper assembly of unfolded polypept...*							●		0.836
tnA-9	*Transposase and inactivated derivative.*	●						●		0.832
kefB	*Glutathione-regulated potassium-efflux system protein; Belongs to the monovalent cation:proton antiporter 2 (CPA2) transporter ...*	●						●		0.831
tnB-9	*Transposase and inactivated derivative; Submitted as non-pseudo.*	●						●		0.826

Your Current Organism:

Orientia tsutsugamushi

NCBI taxonomy Id: 357244

Other names: O. tsutsugamushi str. Boryong, Orientia tsutsugamushi Boryong, Orientia tsutsugamushi str. Boryong, Orientia tsutsugamushi strain Boryong

The predicted partners include histidine kinases(hk25,hk28,and hk45) ,suggesting the protein's role in stress responsive signal transduction pathways.

Step: 6 Biological Interpretation

Done by: Vaidehi Kadam

Presented by :Siddhesh Uday Sapre

Step 6: Biological interpretation

Key findings:

- The protein belongs to the **HtrA/ DegP** family of **serine proteases**.
- It was predicted to be a **periplasmic protein**, not a cytoplasmic one.
- The domain structure showed that the protein is **complete and functional**.
- Overall, the protein was found to be biologically active in the bacteria.

Predicted biological role of TSA47:

- This protein was expected to act as **serine protease**.
- It helps in **removing misfolded** or **damaged proteins**.
- And supports **proper folding** and **maintenance** of **exported proteins**.
- Thus this helps bacteria to **survive under stressful conditions**.

Biological and Pathogenic importance:

- The protein helps *O. tsutsugamushi* to survive inside the host.
- Stress response proteins are important for bacterial survival during infection.
- This protein may help *O. tsutsugamushi* maintain protein stability inside host cells.
- Proper protein maintenance supports bacterial growth and persistence in the host.

Conclusion:

So, based on all the in silico analyses, the protein was predicted to function as a periplasmic HtrA- like serine protease, which is involved in maintaining protein quality and helping manage stress. These functions support bacterial survival during infection. Additionally, TSA47 has been **experimentally shown** to **share homology with human HtrA1**, indicating **conserved structural and functional features** between **bacterial and human serine proteases** (Chen *et al.*, 2009: PMID: 19289508: DOI: 10.1128/IAI.01298-08).

Acknowledgements

Training and Supervision:
Ms. Muskan Kashyap

Team Members:

1.Gayathri Snigdha

2.Paripoorna Reddy

3.Purvi Shah

4.Sameen Khalid

5.Siddhesh Uday Sapre

6.Vaidehi Kadam