

Indian Institute of Technology Gandhinagar



Project-Assignment-1

Name: Nabendu Das

Roll No: 24250059

Branch: Biological Sciences & Engineering

Subject: Biocomputing (BE623)

Submission Date: 19/09/2025

Question 1 (5 Points)

Choose a gene of your interest from human (*Homo sapiens*) and perform the following:

a) Mention the gene name and some description and search for this gene at NCBI.

Refine the query using the methods discussed in the lab sessions and download the gene sequence as fasta file (Take screenshots to show the search) (1 Point)

Gene Name: THAP domain containing 9 (THAP9)

Description: THAP9 is a transposable element-derived gene that encodes the THAP9 protein, which is homologous to the Drosophila P-element transposase and can cut and paste DNA. It enables sequence-specific DNA binding activity and transposase activity.

The screenshot shows the NCBI Gene database page for the gene THAP9. At the top, there is a search bar with "Gene" selected and "THAP9" entered. Below the search bar are links for "Create RSS", "Save search", and "Advanced". The page is set to "Tabular" view, "20 per page", and "Sort by Relevance". The main content area displays the gene name "THAP9 – THAP domain containing 9" and the organism "Homo sapiens (human)". It also lists "Also known as: hTh9" and "Gene ID: 79725". There are buttons for "RefSeq products", "Orthologs", and "Genome Data Viewer". A link "New - Visualize gene across multiple species" is also present. At the bottom, there is a section for "RefSeq Sequences" with a plus sign icon.

The screenshot shows the "Search details" box in the NCBI interface. It contains the search query "THAP9[Gene] AND \"Homo sapiens\"[Organism]". Below the query box is a "Search" button and a "See more..." link.

THAP9 THAP domain containing 9 [*Homo sapiens* (human)]

[Download Datasets](#)

Gene ID: 79725, updated on 19-Aug-2025

Summary

Official Symbol	THAP9 <small>provided by HGNC</small>
Official Full Name	THAP domain containing 9 <small>provided by HGNC</small>
Primary source	HGNC:HGNC:23192
See related	Ensembl:ENSG00000168152 MIM:612537 AllianceGenome:HGNC:23192
Gene type	protein coding
RefSeq status	VALIDATED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	hTh9
Summary	Enables sequence-specific DNA binding activity and transposase activity. Involved in DNA integration and DNA transposition. Predicted to be located in nucleoplasm. [provided by Alliance of Genome Resources, Jul 2025]
Expression	Broad expression in testis (RPKM 2.9), thyroid (RPKM 1.5) and 25 other tissues See more
Orthologs	all
NEW	Try the new Gene table Try the new Transcript table

```
>NC_000004.12:82900735-82919969 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
GGAGCTAAAGTGGTCGTGATTCATGCTGTCGCGGGAACCCCGAAGGTGGGGCCCCACGTAACAAGAAGAT
GACCCGAAGTTGCTCCGAGTGGGCTGCAGCACCCGTGACACCGTGCTCAGCCGGGAGCGCGGCCTCTCC
TTCCACCAGTGCATATGGGAGCAGCCTCGAAGCCTTCGAACCTCCCTGCGGGGCCCCGGCGGGCCGTGGCGT
GGCGTGGGGCGGGGCCGGGCCGCACTGTGGGTGCGCGCGCGTGTGTGACGCGACGTGACGTGCGCAGCGT
CGGGGCGTGGGAGCGGAATTGGGGCACTGTGAGAATTGAGATTCTCTCTCTTCCCGCCAGTGTTTACCT
CTGAATAGCCGGTTCTCGCACCGCCTACCGCTTTAAGGAGAGTTACGCGAAAGGAGTCCTCCATTAATTG
GATGGTCTTATTATTGAAGTGAGGAGGCCGAAAGGGAAGGCAGTTTAGGACAACGGTTGAAGTGTGTGT
GTGGCGTCTTCTGAGCACGAGTACAGATCAGTTTTCTCCTGTACTGCCTGACATTGCAAGTTCTTAAGT
GCTGTTAGCCGAGCTGTGCAAGGCCGGTCCGATTCTGGCTGGATTGATCATGGCAGGGAAGTTTCATTT
GGCTCGACAGTGGGGAAGAATTTAGATCAGTCTTATCTGAACCTCATATCCATGGTTCAGACTGGTATAG
GTCATTAGTATTCAGGGTACACTAGCAAGTAGAAAAGTGGGTAAAATAGATCCTCACAGACAAAAGGAG
CAGATTTATGGAAGATCAAAAAGAACTTTATGGTTGGAAGGGGGGGTAGGAAGAGGGTAGTTAGGAAT
GGCATGGGAAAATGTAAAGGATAGAAAGTGCAATATAGAGGAGTGGAAGGAGTACCTGAGAAACCATTTT
AAAAATGTAAATGCCAGGTTAAGGAGTTAGAGTATGTATGGTAGGCTATGTATAAAATTCATTGATTTT
TTTTTTTTTTTGTCTTATTTTTTAAATAGGGCTGATTCCTCCACCTAAATTGTTAGCAGTGTGAACCA
ACTCATTGTGTATGTGAGACCATAATTGATTTGGGATTTTTCTTGTGATAGAGGTAGTTTTAACTAGGG
AGACATGAATTCCAATTTTAATTGAGTATTTACTGATGTCTGTACTTGGCTCTGGGTCTTGGTTTCTTCT
TCTGTAAGAGGAATAAAACCTGCCTTGTAGAATTTGTAAACTAAAACGTAAGACTTTATAGCATATG
TAACGCACAAGGTAATGCTTTAAAAATGGTATTATTACCAAATTAATTTTTTCCAGAACTCAAACTAG
TTATGCTAGTTAGATTACAATATTTCCCAACATTTTCTTCTTTTTTTTTTTTTTTTTTGGAGACGGG
GTCTCACTTTGTCGCCCAGGCTGGAGAGCAGTGGCACTATCACAGCTCAGCGCAGCCTCTATCTCCAGG
```

b) Find the orthologous (similar) gene from Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Chimpanzee (*Pan troglodytes*) and Fruit fly (*Drosophila melanogaster*) and download the sequences as fasta file. (1 Point)

For Mouse(*Mus musculus*), we didn't find any gene named THAP9. So we searched for a gene named THAP1 in the Mouse, i.e, the orthologous of the THAP9 gene.

Search Query:

Search details

THAP9[Gene] AND "Mus musculus"[Organism]

Search

See more...

Gene

THAP9[Gene] AND "Mus musculus"[Organism]

Save search Advanced

No items found.

Search details

THAP1[Gene] AND "Mus musculus"[Organism]

Search

See more...

Thap1 THAP domain containing, apoptosis associated protein 1 [*Mus musculus* (house mouse)]
Download Datasets

Gene ID: 73754, updated on 30-Jul-2025

Summary

Official Symbol
Thap1 provided by MGI

Official Full Name
THAP domain containing, apoptosis associated protein 1 provided by MGI

Primary source
MGI: MGI:1921004

See related
Ensembl: ENSMUSG00000037214 AllianceGenome: MGI:1921004

Gene type
protein coding

RefSeq status
VALIDATED

Organism
Mus musculus

Lineage
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus

Also known as
4833431A01Rik

Summary
Predicted to enable several functions, including DNA-binding transcription repressor activity, RNA polymerase II-specific; protein homodimerization activity, and zinc ion binding activity. Predicted to be involved in several processes, including DNA-templated transcription; endothelial cell proliferation; and negative regulation of transcription by RNA polymerase II. Predicted to be located in fibrillar center and nucleoplasm. Predicted to be active in nucleus. Is expressed in central nervous system; forelimb bud; and genitourinary system. Used to study torsion dystonia 6. Human ortholog(s) of this gene implicated in torsion dystonia 6. Orthologous to human THAP1 (THAP domain containing 1). [provided by Alliance of Genome Resources, Jul 2025]

Expression
Ubiquitous expression in testis adult (RPKM 5.6), CNS E11.5 (RPKM 3.2) and 28 other tissues See more

Orthologs
human all

```
>NC_000074.7:26648197-26654179 Mus musculus strain C57BL/6J chromosome 8, GRCm39
CGGCGAGCCGAGTCTCGCGAGAGGTGAGGCTGAAGCTGCCTCCGCCATCTTGAAGATGGGAGACGGGCG
ACAGCTGTGGTCCTTCTGCTAAAGCAAACCCACAACGGACAGGGTAGTCACTCCCCACCCCAACCCCA
ACCCACGGCGAGGTGATCGTCCCCGTAAGTGTGACCGACGCCACCGAGAGCGGCGAGCGTTATCAAGG
CCGAGCGCGGCGACCCGACGGCCCCCTTCGCCTGCCTCCCGGGCCGAAGGGAAGTGTGGAGGGCCAGAA
GGATGGTGCAATCCTGCTCCGCCTACGGCTGCAAGAACCCTACGACAAGGACAAGCCCGTCTCCTTCCA
CAAGTAGGTAAGTGGCGTGTCCGCGGCCGCGGCGACAGGGGCGCGCGGGCGGGTGGGGGCGGAGCCCGG
CGCGTGGCCGCGCGAGACCTGGAGGAGGCGGGGCGGGCGCTCGGGCGCTCAGGGCGGCGGACGGGCCGGG
CCGGCGGGGGCGGGGACCCACGGGGGCGCGTCCCCGTGGCTGCGCGCCCCCGCTGCGCATCCCGGGCGCG
CGGCGTCTCCGTTGGTCTTGGCCTCCGCTGGGGGCTGAGACCGCGCGCTTCGCGCTCGCGGAGGGTTTC
CGGCGGGCGGGCAAGCGAGCGAGCGACGCCAGAACCTCTCGCAGCCCTGCGTACCTCTTCGTCCCAAG
CCGCGGCGTGGCCGTGGCACATAGGGAGGAAGTCGAGCTGGGAACGCACAGGGGGAGGAGGACAGGTTTC
GGACCTCACTGTGGATGCCGGAAGGAAGCAGGGCAGCTGACCGGGATGCCAGGGAGGACTCTTAAGTCC
TAGCGTCAAAGAACCAACCTTTAGCATCCATGGGCTTTAGAGGAGGCTCTTGGTATTGTCGAGATGAACC
CCTGTCTTTCAACACACAGGAAGAACGAGACGATAGCACACACTTGACTCTGTTGCCGTGGTGAAGGTTA
ATGTGAGCGAATACCAGCTTTACCCCCACAGCCGTCAGGGCTCGTTCTTGTAGTGGGTGCAGCAATCAAG
AAGCCAGGGTGAAAAATACTCTGGAGAGATGGCTCAGCGGTTGAGAGCACTGACTGCTCTCCAGAGGTC
CTGAGTTCAATTCCCAGCAACCACATGGTGGCTCACAACCATCCGTAATGAGAAATGACGCCTTCTTCTG
ATGTGTCTGAAGACAGCTACAGTGTACTTATATATAATAATAATCTTTTTAAAGAGAAAAATAAAAAAT
ACTCTCTTGAAAGAAGAAAGTCATAGAATTAGCGTTACTGATGGTGTTTGTTAATGCTAACAGTCCACAA
AATGTGACAATTTTTAGGCAGGTAAATTTGTTGTTACAAGTTTTATGTTAGGTTTTGGAGGAATTCCTGA
CTGTATGCAGATGATAGCTAAGAAATCAGATCCACTGGTAAATTATGCAAGTTATCCATAGCAAAACGAT
```

For Rat (*Rattus norvegicus*), we didn't find any gene named THAP9. So we searched for a gene named THAP1 in the Rat, i.e, the orthologous of the THAP9 gene.

Search Query:

Search details

THAP9[Gene] AND "Rattus norvegicus"[Organism]

Search

See more...

Gene

THAP9[Gene] AND "Rattus norvegicus"[Organism]

Save search

Advanced

No items found.

Search details

THAP1[Gene] AND "Rattus norvegicus"[Organism]

Search

See more...

Thap1 THAP domain containing 1 [*Rattus norvegicus* (Norway rat)]

Gene ID: 306547, updated on 5-Sep-2025

Download Datasets

Summary

Official Symbol

Thap1 provided by RGD

Official Full Name

THAP domain containing 1 provided by RGD

Primary source

[RGD:1307589](#)

See related

[Ensembl:ENSRNOG00000056956](#) [AllianceGenome:RGD:1307589](#)

Gene type

protein coding

RefSeq status

PROVISIONAL

Organism

[Rattus norvegicus](#)

Lineage

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Rattus

Summary

Predicted to enable several functions, including DNA-binding transcription repressor activity, RNA polymerase II-specific; protein homodimerization activity; and zinc ion binding activity. Predicted to be involved in several processes, including DNA-templated transcription; endothelial cell proliferation; and negative regulation of transcription by RNA polymerase II. Predicted to be located in fibrillar center and nucleoplasm. Predicted to be active in nucleus. Human ortholog(s) of this gene implicated in torsion dystonia 6. Orthologous to human THAP1 (THAP domain containing 1). [provided by Alliance of Genome Resources, Jul 2025]

Expression

Biased expression in Thymus (RPKM 64.9), Spleen (RPKM 52.6) and 9 other tissues [See more](#)

Orthologs

[human](#) [mouse](#) [all](#)

```
>NC_086034.1:c72612690-72608096 Rattus norvegicus strain BN/NHsdMcwi chromosome 16, GRCr8, whole genome shotgun sequence
GTTGAAGCTGCCTCCGCATCTTGAAGATGGGAGACGAGCGACAGCTGTTGTCTTTCTGCTAAAGCAAAC
ACCACACGGACCGGGTGGTCACCCCCACGGTGAGGTAATCGTCCCATAACTGCTAACCAAAGCCACC
GAAAGCGCGACCGTATCAAGCCGAGCGCGCGGACCGACGGCCCCCTTCACTGCCTCCCGGCGAGAG
GGGAAGTGTGGAGGGCCAGAAGGATGGTGACGCTCTGTCGCTTACGGCTGCAAGAACCAGGTACGATAA
GGACAAGCCCGTCTCTTCCACAAGTAGGTAAC TGCGTGCTCCGCGCGCGGCGAGGGCGCGCGGGCG
GGTGGAGGCGGAGCCCGCGCGTGCCGCGCGAGACCTGGAGGAGCGGGGCGGCGGCGGGCTCGGGCG
GCAGACGGGCTGGCCGCGGGGCGGGGACCATGGCTGCGCGTTCCTCCGCGCGCGCGCTCCGCTTGG
CTGGGCTCTGCTGAGGCTGCGGCGCGGCTTGGCGCTGCGGAGGGTTTCTGGCGAGCAAGCGATGGACG
CCTGGAAGCCTCTCGCTGCCCTCCGGATTTCGTCCCAAGCCAAGCCGTGGAATGTCCCATAGGGATAAA
CTGGAAGTAGAAAAGCATGGGAGGAGACCTGGCTTCTGACCTCGCTGCTGCTGTTAGGAGGGAAGCAGG
ATAGCTGAGTGTGACATCTGACAGGACTTTAGACCTGGTTCAAGAACCAGCTTTAAGGATCTGT
GGGTTTTAGAGTGACTCTTGATTGCAAGATGAGCCCCAGTCTTTCAACACACGGTAGTAATGAAAC
GATAGCATTCACTTTACTCAGATGCCCTGGGAAAGGTTGATGGAACATATTTAGCTTTCCCCACAGCT
GTCAGGCCTTGTCTTGTGTGTCAGCAAGCAAGAATCCAGGATGAAAATACTCTTTGAAAGAAGAAG
GCAATAGAATTAACTTACTGATGGTGATTATTAGCAATGCTAACGGTCCACAAATGTGACATACTTTT
```

For the Chimpanzee (*Pan troglodytes*), we found a gene named THAP9, which is the orthologous of THAP9 in humans.

Search Query:

Search details

THAP9[Gene] AND "Pan troglodytes"[Organism]

Search

See more...

THAP9 THAP domain containing 9 [*Pan troglodytes* (chimpanzee)]

Gene ID: 461209, updated on 10-Jul-2025

Download Datasets

Summary

Official Symbol

THAP9 provided by VGNC

Official Full Name

THAP domain containing 9 provided by VGNC

Primary source

[VGNC:VGNC:729](#)

Locus tag

CK820_G0003631

See related

[EnsemblRapid:ENSPTRG00045002205](#)

Gene type

protein coding

RefSeq status

MODEL

Organism

[Pan troglodytes](#)

Lineage

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Pan

Orthologs

[human](#) [all](#)

```

>NC_072401.2:c47246512-47226708 Pan troglodytes isolate AG18354 chromosome 3, NHGRI_mPanTro3-v2.0_pri, whole genome shotgun sequence
GACAAATAAAGGTCGTAGCCGACAGTCAACGGGCGAGCTAAAGTGGTCGTGATTCATGCTGTCGCGGG
AACCCCGAAGGTGGGGCCACGTAAACAAGAAGATGACCCGAAGTTGCTCCGACGTGGGCTGCAGCACCC
GTGACACCGTCTCAGCCGGGAGCGCGCCTCTCCTTCCACAGTGCATGAGGAGCAGCTCGAAGCCT
TCGAACCTCCCTGCGGGGGCCGGCGGGCGTGGCGCGGCTGGGGCGGGCCGGGCGCACTGTGGGTCGC
GCCGCGTGTGTACGCGACGTGACGTGCGCAGCGTCGGGGCGTGGGAGCGGAATTGGGGCACTGTGAGAA
TTGAGATTCTCTCTCTCCCGCCAGTGTTTACCCTGAGTAGCCGGTCTCGCACCCGCTACCGCTTTA
GGGAGAGTTACGCGAAAGGAGTCCTCCATTAATTGGATGGTCTTATTATTGAAAGTGAGGAGGCCGAAAG
GGAAGGCAAGTTTAGGACAACGGTTGAAGTGTGTGTGGCGTCTCTTGAGCACGAGTACAGATCAGTTT
TCTCTGTACTGCTGACATTGCAAGTTCTTAAGTGTCTTAGCCGAGCTCTGCAAGGCCGGTCCGATTTC
TGGCTGGATTGTATCATGGCAGGGAAGTTTCATTTGGCTCGACAGTGGGGAAGAAATTCAGATCAGTCTT
ATCTGAACCTCATATCCATGGTTTCAGACTGGTATAGGTCAATCAATCAGTATTCAGGGTACACTAGCAAG
TTGAAAGTGGGTAAGTAGATCTCTCAGACAAAGGAGCAGATTTATGGAAGATCAAAAAGAACTTT
ATGGTTGGAAGGGGGGTAGGAAGAGGGTAGTTAGGAATGGCATGGGAAATGTAAAGGATAGAAAGTG
CAAATATAGAGTGGAAAAGTACCTGAGAAACATTTTAAAAATTTAAATGCCAGGTTAAGGAGTTTGAG
TATGTATGGTAGGCTATGTATAAAATTCATTGATTTTTTTTGTTCATTATTTTAAATAGGGCTGATTC
CTCCACCTAAATGTTAGCAGTGTGAACCACTCATTGTGTATGTGAGACATAATTGATTGGGATTTT
TCTTGTGATAGAGGTAGTTTAAACTAGGAGACATGAATCCAAATTTAATTGAGTATTTACTGATGTC
TGACTTGGCTCTGGGCTTGGTTCTTCTCTGTAAAGAGGAATAAAACCTGCCTTGTAGAATTTGTA
AAACTAAACGTAAGACTTTATAGCATGTAGCACACAAGGTAATGCTTTAAAAATGGTATTATTACC

```

For the Fruit fly (*Drosophila melanogaster*), we didn't find any gene named THAP9. So we searched for a gene named piwi (P-element transposases) in the Fruit Fly, i.e, orthologous to THAP9 in humans.

Search Query:

Search details

THAP9[Gene] AND "Drosophila melanogaster"[Organism]

Search

See more...

Gene THAP9[Gene] AND "Drosophila melanogaster"[Organism]

Save search Advanced

No items found.

Search details

P-element[All Fields] AND transposase[All Fields] AND "Drosophila melanogaster"[Organism]

Search

See more...

piwi P-element induced wimpy testis [<i>Drosophila melanogaster</i> (fruit fly)]		Download Datasets
Gene ID: 34521, updated on 17-Aug-2025		
Summary		
Official Symbol	piwi provided by FlyBase	
Official Full Name	P-element induced wimpy testis provided by FlyBase	
Primary source	FLYBASE:FBgn0004872	
Locus tag	Dmel_CG6122	
See related	AllianceGenome:FB:FBgn0004872	
Gene type	protein coding	
RefSeq status	REVIEWED	
Organism	<i>Drosophila melanogaster</i>	
Lineage	Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora	
Also known as	CG6122; DmelCG6122; Piwi; PIWI	
Summary	Enables RNA endonuclease activity; chromatin DNA binding activity; and piRNA binding activity. Involved in several processes, including germ cell development; pole cell formation; and regulatory ncRNA-mediated gene silencing. Located in several cellular components, including chromatin; chromocenter; and nucleus. Part of RNA polymerase II transcription repressor complex. Is active in cytoplasmic ribonucleoprotein granule; cytosol; and nucleus. Is expressed in several structures, including adult brain; adult fat body; germline cell; gonad; and somatic cell of testis. Used to study primary ovarian insufficiency. Orthologous to human PIWIL1 (piwi like RNA-mediated gene silencing 1); PIWIL3 (piwi like RNA-mediated gene silencing 3); and PIWIL4 (piwi like RNA-mediated gene silencing 4). [provided by Alliance of Genome Resources, Jul 2025]	
Orthologs	all	


```
>NT_033779.5:c10987420-10982205 Drosophila melanogaster chromosome 2L
TCACACTGAGTCCAAAGCGTCGTTTTCAAAGTACTCTTTCAGTTTCCATTGTGAAGTTTTAAGTGATCGC
GAGTGCCAAAAAGTAACAATGGCTGATGATCAGGGACGTGGACGCAGGCGTCCACTTAACGAAGATGATT
CCTCTACTTCCCGAGGTAGTGGTGATGGGCGGTATGTAAACTCTTATTCAAGCTAGGAAAAAAATCAT
AGCGGTAAAGCAAATTCCTTATAATTCCGTTATGTCGCAAACACAAACACAGCCATGTACATAGATTGTA
AATGCGCGCATTCTTTCGGAATTCGACACCAAAAGGAAGCACTAAAACTAACAACAAAAATCTGCTATG
AATTTGAGCTTGACTGCATAAAATCTTCTTGAATTTCTGTTGACCAAGCACTTTTTTCAAAGGGAATTCA
TTAAACATATTTCCAAAAACGCTTCTTCAATTTGAATTTCCAAGGAAATGTAAAGCATTATCTGTTTCG
TTTCATTTATTTTTTTTTTAATAAAAACTGCTAAAATAACCATCACCTGACAGGAAGGTGGCTGTGGCAAG
CCGAATCTTATTTACCTTGCAGTCTATGTCTTTGGAATCTGCATGCCCCGACGTTGATATGGACGGACAT
GACCAAATGGACCCGGCTAGTGATACTCATAAAAATCATGGTCGGAACGCTTCCTTCTATCTGTCACA
TACATTTGATTCTAGAACTCTACGAGTAACGGGTATTTCTCCGATATTCATCCCAATTCTATACAAC
TTGAGATGGCTTTCAGCATTGATATACTCTGAAATTGCAGGCAAAGCATACTACAAAACTATTTTTTAC
CATTTTTTTTCTGCTTCATCTTTGATTATGAGTAACAATATTTTCGTAGCTTGACCGTTAGTTTTATGAT
TTAAAATCTGTCGTACTTTAGGTAGTTGTTTAAATGGCTCACATTTAGGAATCACAACATTTTGTTAAA
CCCCTTAATATTTTTTCTTTTAATAATAAATAATAAACAGGTAATACTTTTTGATTGTTTCATCGCGATTG
TTTTCAATGTACAAAGTTGAGTCTTTCTGGTGTTTATTGCTGTCTCGTGTTTATTGCTGAATCTCGCACG
ATCTCTCTCTTTCTTCTATGTTATTTTTTCCATCGGCGTCAATTATTCATAACTGCCGCTAGGTGGAGCT
AGTGTGCAGTCATTCACTTCTATTAGGAGACAGCGTCTCTATATCTAATATTATAATTTTTTTTAAATG
AGTAACCTACATCATTTAATGCTTTCATGCAGCGGTGAAAGTATTCAGAGGATCTTCATCAGGTGACCC
GAGAGCGGATCCTCGTATAGAGGCTTCAAGAGAGAGAAGAGCTCTCGAGGAAGCTCCCAGGCGTGAAGGT
GGCCCCGACAGAGCGAAAAGCCGTGGGGTGACCAATATGATTACCTGAATACCCGTCCGGCTGAGCTGGTAT
CCAAGAAGGGAACCGATGGCGTCCCGTCTATGCTGCAGACGAACTTTTTCCGATTAAAAACCAAGCCGGA
```

c) For each of these five sequences, extract the header and save in a separate file (1 Point)

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed -n '/^>/p' THAP9_human.fasta THAP9_chim
panzee.fasta THAP1_rat.fasta THAP1_mouse.fasta piwi_fruit_fly.fasta > header_files.txt
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ vi header_files.txt
```

```
>NC_000004.12:82900735-82919969 Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly
>NC_072401.2:c47246512-47226708 Pan troglodytes isolate AG18354 chromosome 3, NHGRI_mPanTro3-v2.0_pri, whole genome shot
gun sequence
>NC_086034.1:c72612690-72608096 Rattus norvegicus strain BN/NHsdMcwi chromosome 16, GRCr8, whole genome shotgun sequence
>NC_000074.7:26648197-26654179 Mus musculus strain C57BL/6J chromosome 8, GRCh39
>NT_033779.5:c10987420-10982205 Drosophila melanogaster chromosome 2L
```

So in this step, we extract five header lines and save them to a .txt file named header_files.txt

d) calculate the GC content of each sequence (2 Points)

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' THAP9_human.fasta | awk '{gc+=g
sub(/[GC]/," "); total+=length($0)} END {print "GC% =", (gc/total)*100}'
GC% = 36.9743
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' THAP9_chimpanzee.fasta | awk '{
gc+=gsub(/[GC]/," "); total+=length($0)} END {print "GC% =", (gc/total)*100}'
GC% = 36.8543
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' THAP1_rat.fasta | awk '{gc+=gsu
b(/[GC]/," "); total+=length($0)} END {print "GC% =", (gc/total)*100}'
GC% = 44.0261
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' THAP1_mouse.fasta | awk '{gc+=g
sub(/[GC]/," "); total+=length($0)} END {print "GC% =", (gc/total)*100}'
GC% = 45.3953
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' piwi_fruit_fly.fasta | awk '{gc
+=gsub(/[GC]/," "); total+=length($0)} END {print "GC% =", (gc/total)*100}'
GC% = 39.6281
```

Here, I have calculated the GC content of each sequence using sed and awk commands.

Gene Name	% GC content
THAP9 (Human)	36.97
THAP9 (Chimpanzee)	36.85
THAP1 (Rat)	44.03
THAP1 (Mouse)	45.40
piwi (Fruit Fly)	39.63

Question 2 (15 Points)

For the same chosen gene:

a. Use the coding sequence (CDS) of this gene to translate to amino acid sequence. You may reduce the gene length to only 30 nucleotides for ease. (Hint: use Genetic code table and search/replace, Coding sequence can be downloaded from NCBI) (3 points)

Coding sequence:

```
>NM_001317776.2:567-2846 THAP9 [organism=Homo sapiens] [GeneID=79725] [transcript=5] [region=cds]
ATGTTACAAGTGTCACAAAAAAGACTTATCTCCGTAAAGAACTACAGGATGATCAAGAAGAGAAAGGGTT
TACGATTAATTGATGCACTTGTAGAAGAGAACTACTTTCTGAAGAAACAGAGTGCTGCTACGAGCTCA
ATTTTCAGATTTTAAGTGGGAGTTATATAATTGGAGAGAAACAGATGAGTACTCCGCAGAAATGAAACAA
TTTGATGTACACTCTACTTGTGCAAGTCTATGATTATGTAAGAAAGATTCTTAAGCTGCCTC
ATTCTTCCATCCTCAGAACGTGGTTATCCAAATGCCAACCCAGTCCAGGTTTCAACAGCAACATTTTTC
TTTTCTTCAACGAAGAGTAGAGAATGGAGATCAGCTCTATCAATACTGTTTCATTGTTAATAAAAAGTATG
CCTCTCAAGCAACAGCTTCAGTGGGATCCTAGCAGTCACAGTTTGCAGGGGTTTATGGACTTTGGTCTTG
GAAAACTTGATGCTGATGAAACGCCACTTGCTTCAGAAACTGTTTTGTAAATGGCAGTGGGTATTTTGG
CCATTGGAGAACACCTCTTGTTATTTTTTTGTAAACAGAGCATCTGGATATTTGCAGGCTCAGCTGCTT
CGTCTGACTATTGGTAACTGAGTGACATAGGAATCACAGTTCTGGCTGTTACATCTGATGCCACAGCAC
ATAGTGTTTCAGATGGCAAAAGCATTGGGGATACATATTGATGGAGACGACATGAAATGTACATTTTCAGCA
TCCTTCATCTTCTAGTCAACAGATTGCATACTTCTTTGACTCTTGCCACTTGCTAAGATTAATAAGAAAT
GCATTTCAGAATTTTCAAAGCATTTCAGTTTATTAATGGTATAGCACATTGGCAGCACCTCGTGGAGTTAG
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' cds.fna | tr -d '\n' | cut -c1-30
ATGTTACAAGTGTCACAAAAAAGACTTATC
```

So I got this sequence,

ATGTTACAAGTGTCACAAAAAAGACTTATC

To translate it into an amino acid sequence, I use EMBOSS Transeq. It uses the standard genetic code to code the amino acids.

EMBOSS Transeq

Sequence Translation (ST)

[Job Dispatcher](#) [Help & Privacy](#) [Your Jobs](#) **Input form** [Feedback](#)

Welcome to the Job Dispatcher website! If you need assistance or have feedback, please [contact](#) us.

EMBOSS Transeq translates nucleic acid sequences to their corresponding peptide sequences. It can translate to the three forward and three reverse frames, and output multiple frame translations at once.

Input sequence ⓘ

Paste your sequence here - or use the example sequence

ATGTTACAAGTGTCCAAAAAAGACTTATC

Choose File

No file chosen

Use the example

Clear sequence

More example inputs

Parameters

FRAME ⓘ

1

▼

CODON TABLE ⓘ

Standard Code

▼

More options ▼

Tool output

>EMBOSS_001_1

MLQVSKKRLI

Download

Alignment with colours

>EMBOSS_001_1

MLQVSKKRLI

Hide

So we got these amino acid sequences: MLQVSKKRLI, from EMBOSS Transeq as an output.

b. Retrieve the protein sequence for this gene from UniProt in FASTA format and compare the sequence of gene translated by you with that retrieved from Uniprot. (2 points) (Hint: You may use bash commands to compare files or strings)



UniProtKB BLAST Align Peptide search ID mapping SPARQL Release 2025_03 | Statistics

Find your protein




UniProtKB THAP9 X Advanced | List Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

Q9H5L6 · THAP9_HUMAN

Protein ⁱ	DNA transposase THAP9	Amino acids	903 (go to sequence)
Gene ⁱ	THAP9	Protein existence ⁱ	Evidence at protein level
Status ⁱ	 UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	 (4/5)
Organism ⁱ	Homo sapiens (Human)		

Entry Variant viewer 849 Feature viewer Genomic coordinates Publications External links History

 Tools  Download  Add Add a publication Entry feedback

```
<  >  ↺  rest.uniprot.org/uniprotkb/Q9H5L6.fasta

>sp|Q9H5L6|THAP9_HUMAN DNA transposase THAP9 OS=Homo sapiens OX=9606 GN=THAP9 PE=1 SV=2
MTRSCSAVGCSTRDTVLSRERGLSFHQFPTDTIQRSKWIRAVNRVDPRSKKIWIIPGPGAI
LCSKHFQESDFESYGIRRLKKGAVPSVSLYKIPQGVHLKGKARQKILKQPLPDNSQEVA
TEDHNYSLKTPLTIGAELAEVQQLQVSKKRLISVKNYRMKKRGLRLIDALVEEKL
SEETECLLRAQFSDFKWEIYNWRETDEYSAEMKQFACTLYLCSSKVYDYVRKILKLP
HSSILRTWLSKQPSPGFNSNIFSLQRRVENGQQLYQYCSLLIKSMPLKQQLQWDPSSHSLQ
GFMDFGLGKLDADETPLASETVLLMAVGIFGHWRTPLGYFFVNRASGYLQAQLRLTIGK
LSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQHPSSSSQQIAYFFDSCHLLR
LIRNAFQNFQSIQFINGIAHWQHLVELVAEEQELSNMERIPSTLANLKNHVLKVNATQ
LFSESVASALEYLLSLDLPPFQNCIGTIHFLRLINLFDIFNSRNCYKGLKGPLLPETY
SKINHVLIEAKTIFVTLSDTSNQTIKQKQLGFLGFLNAESLKWLYQNYVFPKVMFPF
YLLTYKFSHDHLELFLKMLRQVLVTSSSPTCMAFQKAYNLETRYKFQDEVFLSKVSIFD
ISIARRKDLALWTQRYQYGVSVTKTVFHEEGICQDWSHCSEALDLSDHRRNLICYAG
YVANKLSALLTCEDCITALYASDLKASKIGSLLFVKKKNGLHFPSESLCRVINICERVVR
THSRMAIFELVSKQRELYLQKILCELSGHINLFVDVNKHLFDGEVCAINHFVKLLKDII
ICFLNIRAKNVAQNPLKHHSERTDMKTLSRKHWSVQDYKCSSFANTSSKFRHLLSNDGY
PFK
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ wget -O THAP9_uniprot.fasta https://rest.uniprot.org/uniprotkb/Q9H5L6.fasta
--2025-09-17 12:30:07-- https://rest.uniprot.org/uniprotkb/Q9H5L6.fasta
Resolving rest.uniprot.org (rest.uniprot.org)... 193.62.193.81
Connecting to rest.uniprot.org (rest.uniprot.org)|193.62.193.81|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1007 [text/plain]
Saving to: 'THAP9_uniprot.fasta'

THAP9_uniprot.fasta      100%[=====] 1007 --.-KB/s   in 0s

2025-09-17 12:30:08 (67.1 MB/s) - 'THAP9_uniprot.fasta' saved [1007/1007]

(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ sed '/^>/d' THAP9_uniprot.fasta | tr -d '\n' | cut -c1-10
MTRSCSAVGC
```

So from Uniprot, the first 10 amino acid sequences are MTRSCSAVGC. So it is not matched with the sequence translated by me. I think the probable reason may be the differences in the gene annotation format between NCBI and Uniprot.

c. Use the protein sequence from Uniprot to run a BLASTP search at NCBI and download the top 5 homologous sequences as a FASTA file. (1 Point)

blastn
blastp
blastx
tblastn
tblastx

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

>sp|Q9H5L6|THAP9_HUMAN DNA transposase THAP9 OS=Homo sapiens
OX=9606 GN=THAP9 PE=1 SV=2
MTRSCSAVGCSTRDTVLSRERGLSFHQFPTDTIQRSKWIRAVNRVDPRSKKI
WIPGPGAI

Query subrange [?](#)
From
To

Or, upload file No file chosen [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Job Title **sp|Q9H5L6|THAP9_HUMAN DNA transposase THAP9...**
RID [CMS3VTKG014](#) Search expires on 09-18 15:17 pm [Download All](#) [v](#)
Program BLASTP [?](#) [Citation](#) [v](#)
Database **ClusteredNR** [See details](#) [v](#)
Query ID lcl|Query_11856818
Description sp|Q9H5L6|THAP9_HUMAN DNA transposase THAP9 OS= ...
Molecule type amino acid
Query Length 903
Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear **NEW**

[+ Add organism](#)

Percent Identity to
E value to
Query Coverage to

Clusters
Graphic Summary
Alignments
Taxonomy

	Cluster Composition	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
v	38 member(s), 27 organism(s)	placentals	DNA transposase THAP9 isoform 1 [Homo sapiens]	1873	1873	100%	0.0	100.00%	903	NP_078948.3
v	7 member(s), 5 organism(s)	rodents	DNA transposase THAP9 [Marmota flaviventris]	1674	1674	100%	0.0	88.93%	901	XP_027795633.1
v	5 member(s), 2 organism(s)	armadillos and others	DNA transposase THAP9 isoform X1 [Dasypus novemcinctus]	1662	1662	100%	0.0	88.95%	900	XP_058152454.1
v	93 member(s), 63 organism(s)	placentals	Dna Transposase Thap9 [Manis pentadactyla]	1656	1656	98%	0.0	89.86%	943	KAI5278805.1
v	1 member(s), 1 organism(s)	Bolivian squirrel monkey	DNA transposase THAP9 [Saimiri boliviensis]	1656	1656	100%	0.0	90.37%	999	XP_074251769.1

So I have downloaded the top 5 homologous sequences from running a blastp of THAP9.

d. From this combined FASTA file, extract all the sequence IDs and save them in a separate file. (1 point)

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^>/ {print substr($1,2)}' THAP9_homologs_5seq.txt > sequence_id5.txt
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ vi sequence_id5.txt
```

```
NP_078948.3
XP_027795633.1
XP_058152454.1
KAI5278805.1
XP_074251769.1
```

Here, first I print all the sequence IDs using awk, and then save them to a separate file named sequence_id5.txt.

e. Extract sequences that contain a hydrophobic residue followed by a charged residue followed by an aromatic residue and save them in a separate FASTA file. (2 points)

So hydrophobic residues in the protein sequence are Alanine(A), Valine(V), Isoleucine(I), Leucine(L), Methionine(M), Phenylalanine(F), Tyrosine(Y), and Tryptophan(W).

Charge residues are Arginine (R), Histidine(H), Lysine(K), Aspartic Acid(D), and Glutamic Acid(E).

Aromatic residues are Phenylalanine(F), Tryptophan(W), Tyrosine(Y).

Chatgpt prompt:

e. Extract sequences that contain a hydrophobic residue followed by a charged residue followed by an aromatic residue and save them in a separate FASTA file. How to proceed

Chatgpt result:

3. Using `grep` (if sequences are on one line each)

bash

Copy code

```
grep -B1 -E '[AVILMFYW][KRHDE][FWY]' THAP9_homologs_5seq.txt > THAP9_pattern_hits.fasta
```

- `-E` → regex mode.
- `[AVILMFYW][KRHDE][FWY]` → hydrophobic-charged-aromatic motif.
- `-B1` → include 1 line before the match (the FASTA header).

Using Chatgpt I know the use of -E, -B1, and how I write a hydrophobic-charged-aromatic motif. So, find these motifs using grep search.

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ grep -B1 -E '[AVILMFVY][KRHDE][FWY]' THAP9_homoLogs_5seq.txt > THAP9_hca.txt
```

```
KKGAVPSVSLYKIPQGVHLKGKARQKILKQPLPDNSQEVATEDHNYSLKTPLTIGAEKLAEVQQMLQVSKKRLISVKNYR
MIKKRKGLRLIDALVEEKLLSEETECLLRAQFSDFKWELYNWRETDEYSAEMKQFACTLYLCSSKVYDYVRKILKLPHSS
ILRTWLSKCQSPGFSNIFSFLQRRVENDQLYQYCSLLIKSMPLKQQLQWDPSSHSLQGFMDFFGLGKLDADETPLASE
TVLLMAVGIFGHWRTPLGYFFVNRASGYLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQ
HPSSSSQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVALEEELSNMERIPSTLANLKNHVLKVN SATQ
LFSESVASALEYLLSLDLPPFQNCIGTIFHLRLINNLFDFNSRNCYGKGLKGPLL PETYSKINHVLEAKTIFVTLSDT
SNNQIIKGKQKLGFLGFLNAESLKWLYQNYVFPKVPFPYLLTYKFSQDHLELFLKMLRQVLVTSSSPTCMFQKAYYN
LETRYKQDEVFLSKVSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDWSHCSLSEALLDLSDHRRNLICYAG
YVANKLSALLTCEDCITALYASDLKASKIGSLLFVKKKNGLHFPSESLCRVINICERVVRTHSRMAIFELVSKQRELYLQ
```

```
KKGAVPSVSLYKIPQGVHLKGKARQKILKQPLPDNSQEVASEDHNYSLRRPWTGGAEKLAEVQQMLQVSKKRLISVKNYR
IIKKRKGLQLIDSLVEERLLSEETECLLRAQFSDFNWEMYNWRETAEYSTEMKQFACTLYLCSSKVYDYVRKILKLPHSS
ILRTWLSRCKPSPGFSNIFSFLQRRVENDQIYQYCSLLIKGISLKEQLQWDPSSNSLQGFMDFFGLGKLDADETPLASE
TILLMAVGIFGHWRTPLGYFFVNRSGYLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQ
HPSSSSQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVALGEQELSYMKGIPRKHANLKNHVLKMNCAAQ
LFSESVASALEYLLSVGLPPFQNCIGTIFHLRLINNLFDFNSRNCYGKGLKGPLL PETFSKINHVLEAKTIFLTLSDT
SNNQILKGKRLGFLGFLNAESLKWLYQNYVFPKVPFPYLLTYKFSQDHLELFLKMLRQVLVPSSSFTCMFQKAYHN
LETRYRFQDEVFLSEVSIFDISVARRKDLALWTVQRQYGVVRMKSFLHKEDICQDWSNCSLSEALLDLSQKRS LTCYAG
YIATRLSAILTCEDCISALYASDLKASKIGSLLCVKKKNGLHFPSESLCQIINICEQVLRTHSRMVVFELVPKHRELYLQ
```

```
MTRSCSAVGCSTRDVL SRERGLSFHQCVFPTDTIQRAKWIRAVNRVDPRSKKIWIPIPGGAILCSKHFQESDFESYGIRR
KLKKGAVPSVSLYKVLQAADLKGKARQKILKQPLPDNSQEVATEDHNYSLKTPLIIGAEKLAEVQQMLQVSKKRLISVKN
YRLIKKKRGLRLIDALVEEKLLSEETELLRAQFSDFKWELYNWRETAEYSTEMKQFACTLYLCSSKVYDYVRKILKLPH
SSILRTWLSKCQSPGFSNIFSFLQRRVENDQLYQYCSLLIKGISLKEQLQWDPSSHRLQGFMDFFGLGKLDADETPLA
SETILLMAVGIFGHWRTPLGYFFVNRSSGYLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIRIDGDNVKT
FQHPSSSSQIAYFFDSCHLLRLIRNAFYFQSIQFINGIAHWQHLVELVALKEQELSSTERISSKFKPKLKNHVLKMNCA
AQLFSESVASALECLLSLGLPPFQNCIGTIFHLRLINNLFDFNSRNYGKGLKGPLL PETYNKINHVLEAKTIFVTLS
DTSNNQIIKGKRLGFLGFLNAESLKWLYQNYVFPKVPFPYLLTYKFSQDHLELFLKMLRQVLVSSSPTCMFQKAY
HNLETRYRFQDKAFLEVSILDISIARRKDLALWTIQHEYGGSIIKALFHKEVICQDWSNCSLSEALLDLSDHRRNL TTY
AGYIATKL TALLTCEECITALYASDLKAFKIGSLLCVKKKNGLHFPSESLCRIINICERVVRTHSRMGIHELVPKQMDLY
```

```
KKGAVPSVSLYKVLQGVHLKGKARQKILKQPLPDNSPEFATEDHNYSLKRPLTIGAEKLAEVQQMLQVSKKRLISVKNYR
MIKKRKGLRLIDALVEEKLLSEETECLLRAQFSDFKWELYNWRETAEYSTEMKQFACTLYLCSSKVYDYVRKILKLPHSS
ILRTWLSKCQSPGFSNIFSFLQRRVENDQLYQYCSLLIKGISLKEQLQWDPSSHRLQGFMDFFGLGKLDADEMPLASE
TILLMAVGVSQHWRTPLGYFFVNRASGYLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQ
HPSSSSQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVALEEELANMERIPRKLANLKNHVLKMNFAAQ
LFSESVASALECLLSLGLPPFQNCIGTIFHLRLINNLFDFNSRNCYGKGLKGPLL PETFSKINRVLEAKTIFVTLSDT
SNNQIIKGKRLGFLGFLNAESLKWLYQNYVFPKVLFPYLLTYKFSQDHLELFLRLLRQVLVNSSNPTCLAFQKAYHN
LETRYRLQDETFLSEVSILDISIARRKDLALRTVQRQYDVSIKTLFHKEDFCQDWSNCSLSEALLDLSDHMRNLNCCAG
YIANKLSALLTCEDCISALYASDLRASKIGSLLCVKKKNGLHFPSESLCRVINICERVLRTHSRMAVYELLPKQRELYLQ
```

```
MVGRELPPYFCIGKLQDKAQEHDTLKGGA VKYTSSGLGPSDWLLRQNVACDWLQCRRAFGEEVWESRASKARLSCQIK
PLTVGAEKLAEVQQMLQVSKKRLISVKNYKMIKKRKGLPLIDALVEEKLLSEETECLLRAQFSGAASEAGGKPGDCAGLE
AKCRKYFKEGVVHYVKCCWEIRWLSKCQPGPGFSNIFSFLQRRVQNGDQLYQYCSLLIKSIPLKQQLQWDPSSHSLQGF
MDFFGLGKLDADETPLATETVLLMAVGVFQHWRTPLGYFFVNRASGCLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQM
AKALGIHIDGDDMKCTFQHPSSCSQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVALEEELSNMERIP
STPANLKNHILKVNCAAQLFSESVASALEYLLSLGLPPFQNCIGTIFHLRLINNLFDFNSRNCYGKGLKGPLL PETYSK
INHVLEAKTIFVTLSSSNNQIIQKQKLGFLGFLNAESLKWLYQNYVFPKVPFPYLLTYKFSQDHLELFLKMLRQV
LVTSSSPTCMFQKAYHNLETRCTFQDEVFLSKVSIFDISVARRKDLALWAVQRQYGVSVTKTLFHKEGICQDWSNCSLS
EALLDLSNRRNLICYAGYVADKLSALLTCEDCITALYASDLKASKIGSLLFVKKKNGLHFPSESLCQVINICERVLRTH
```

Here, I searched the motifs using grep search in the entire THAP9 5 homologs sequence and saved it to a text file named THAP9_hca.txt.

f. Calculate the percentage of cysteine residues in each sequence and report it. (2 points)

Chatgpt Prompt:

I have a file that have multiple fasta sequence in a .txt file, So i need percentage of cystine residue in each organism sequence, how I proceed

Great question, Nabendu! Since your `.txt` file contains multiple FASTA sequences, you can calculate the **percentage of cysteine (C) residues** in each sequence using a combination of `awk`, `sed`, and basic shell commands. Here's a simple and clear way to proceed:

Step-by-Step Bash Workflow

- ✓ 1. Flatten the FASTA file (combine multi-line sequences)

This ensures each sequence is on a single line after its header:

Chatgpt Result:

```
Bash Copy  
awk '/^>/ {if (seq) print seq; print; seq=""; next} {seq=seq $0} END {print s
```

Replace `your_file.txt` with the actual filename.

- ✓ 2. Calculate % of cysteine residues per sequence

```
Bash Copy  
awk 'BEGIN{RS=">"; ORS=""} NR>1 {  
  header = substr($0, 1, index($0, "\n") - 1);  
  seq = substr($0, index($0, "\n") + 1);  
  total = length(seq);  
  c_count = gsub(/C/, "", seq);  
  percent = (c_count / total) * 100;  
  printf "%s: %.2f%% cysteine\n", header, percent;  
}' flat.fasta
```

What This Does

- `RS=">":` Treats each FASTA entry as a record.
- `substr(...):` Separates header and sequence.
- `length(seq):` Gets total amino acids.
- `gsub(/C/, "", seq):` Counts cysteine residues.
- `%.2f%%:` Formats the percentage to two decimal places.

In this question, I use chatgpt to understand how to proceed with the question step by step. So, for this, I know about the RS, substr(), and other things that I knew before.

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^>/ {if (seq) print seq; print; seq=""
"; next} {seq=seq $0} END {print seq}' THAP9_homologs_5seq.txt > THAP9_online.fasta
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk 'BEGIN{RS=">"; ORS="" } NR>1 {header = s
ubstr($0, 1, index($0, "\n") - 1); seq = substr($0, index($0, "\n") + 1); total = length(seq); cys_count = gsub(/C/, "",
seq); percent = (cys_count / total) * 100; printf "%s: %.2f%% cysteine\n", header, percent;}' THAP9_online.fasta
NP_078948.3 DNA transposase THAP9 isoform 1 [Homo sapiens]: 2.65% cysteine
XP_027795633.1 DNA transposase THAP9 [Marmota flaviventris]: 2.77% cysteine
XP_058152454.1 DNA transposase THAP9 isoform X1 [Dasypus novemcinctus]: 2.66% cysteine
KAI5278805.1 Dna Transposase Thap9 [Manis pentadactyla]: 2.75% cysteine
XP_074251769.1 DNA transposase THAP9 [Saimiri boliviensis]: 3.40% cysteine
```

Sequence Header Name	% Cysteine Residue
NP_078948.3 DNA transposase THAP9 isoform 1 [Homo sapiens]	2.65
XP_027795633.1 DNA transposase THAP9 [Marmota flaviventris]	2.77
XP_058152454.1 DNA transposase THAP9 isoform X1 [Dasypus novemcinctus]	2.66
KAI5278805.1 Dna Transposase Thap9 [Manis pentadactyla]	2.75
XP_074251769.1 DNA transposase THAP9 [Saimiri boliviensis]	3.40

In this question, I have calculated the percentage of cystine residue in each sequence using the awk command and printed the header-wise cysteine percentage in each sequence.

g. From the BLAST results, find the longest sequence and report its ID and its length. For this sequence, calculate the amino acid frequency and provide a distribution of each residue type. (4 points)

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^>/ {if (len > max) {max=len; id=hdr}
hdr=$0; len=0} /^[^>]/ {len+=length($0)} END {if (len > max) {max=len; id=hdr} print id, max}' THAP9_homologs_5seq.txt
>XP_074251769.1 DNA transposase THAP9 [Saimiri boliviensis] 999
```

First, I find the longest sequence and report its ID and length using awk command. So I find
 >XP_074251769.1 DNA transposase THAP9 [Saimiri boliviensis] sequence length is higher i.e. 999.

Chatgpt Prompt:

How I calculate amino acid frequency's of the longest sequence, solve using linux

Perfect! Let's do this entirely in Linux using command-line tools. I'll assume your file is a multi-FASTA file called `sequences.fasta`.

Chatgpt Result:

bash

Copy code

```
awk '/^>/{if(seq!=""){print header"\n"seq}; header=$0; seq=""}/^[^>]/ {seq=seq $0} END{print header"\n"seq}' THAP9_homologs_5seq.txt
```

bash

Copy code

```
grep -v '^>' longest_seq.fasta | fold -w1 | sort | uniq -c | awk '{printf "%s\t%s\t%.2f%%\n", $2, $1, ($1/999)*100}' | sort
```

So, using chatgpt I learn how to calculate the amino acid frequency of the longest sequence. Then `grep -v` removes the header lines and leaves only the sequence lines, `fold -w1` wraps the sequence to one character per line, `sort` is used to sort the single-letter lines alphabetically so identical amino acids group together, `uniq -c` counts how many times each unique letter appears and prints lines.

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^>/{if(seq!=""){print header"\n"seq}; header=$0; seq=""}/^[^>]/ {seq=seq $0} END{print header"\n"seq}' THAP9_homologs_5seq.txt | awk '/^>XP_074251769.1/{print;getline;print}'
>XP_074251769.1 DNA transposase THAP9 [Saimiri boliviensis]
MVGRELPPYFCIGKLQDKAQEHDTLKGAVKYTSSGLGPSDWLLRQNVACDWLQCRRAFGEEVWESRASKARLSCQIKVAAEAESTGGAKVIVIPAFRGPEGDPDPRNEKMTTRSCSAVGC
STRDVLRSERGLSFHQFPDDTIQRSKWIRAVNRVDPKSKKIWIPGPGAILCSKHFQESDFESYGIRRLKKGTVPSVSLYKVPQGVHLKGARQKILKQPLPINSQEVATEDHNYSLKT
PLTVGAEKLAEVQQLQVSKRLISVKNYKMIKKRGLPLIDALVEEKLLSEETECLLRAQFSGAASEAGGKPGDCAGLEAKCRKYFKEGVVHYVKCCWEIRWLSKCQPGPGFNHSIFSF
LQRRVQNGDQLYQYCSLIISIPKQQLQWDPSSHSLQGFMDFGLGKLDADETPLATETVLLMAVGVFQHWRTPLGYFFVNRASGCLQAQLRLRTIGKLSDIGITVLAVTSDATAHSVQM
AKALGIHIDGDDMKCTFQHPSSCSQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVALEEQLSNMERIPSTPANLKNHILKVNCAAQLFSESVASALEYLLSLGLPPFQ
NCIGTTHFLRLINNLFDIFNSRNCYKGLKGPLLPETYSKINHVLTEAKTIFVTLSDDSSNNQIIQGGKQLGFLGFLNAESLKWLYQNYVFPKVPFPYLLTYKFSHDHLEFLKMLRQV
LVTSSSPTCMFAQKAYHMLETRCTFQDEVFLSKVSIFDISVARRKDLALWAVQRQYGVSVTKTLFHKEGICQDWSNCSLSEALLDLSDNRRNLICYAGYVADKLSALLTCEDCITALYAS
DLKASKIGSLLFVKKKNGLHFPSESLCQVINICERVLRTHSRMAIFELIPKQRELYLQKILCELSGHIYLFVDLKNHILFDGEVCAINHVFVKLLKDIIICFLNIRANNVAQNPLKHSER
DMKTLRKHWSLDPDYKCSFANTNKRHLLSNDGYPEK
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^>/{if(seq!=""){print header"\n"seq}; header=$0; seq=""}/^[^>]/ {seq=seq $0} END{print header"\n"seq}' THAP9_homologs_5seq.txt | awk '/^>XP_074251769.1/{print;getline;print}' > longest_seq.fasta
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ grep -v '^>' longest_seq.fasta | fold -w1 |
sort | uniq -c | awk '{printf "%s\t%s\t%.2f%%\n", $2, $1, ($1/999)*100}' | sort
A      63      6.31%
C      34      3.40%
D      41      4.10%
E      49      4.90%
F      50      5.01%
G      59      5.91%
H      31      3.10%
I      58      5.81%
K      77      7.71%
L     121     12.11%
M      14      1.40%
N      41      4.10%
P      39      3.90%
Q      53      5.31%
R      49      4.90%
S      82      8.21%
T      40      4.00%
V      56      5.61%
W      14      1.40%
Y      28      2.80%
```

Question 3 (7 Points)

a. Retrieve the 3D structure of the same protein from the RCSB PDB database in PDB format. If the PDB structure is not present, then download the PDB file of the AlphaFold model from AlphaFold database. (1 Points)

I don't find the THAP9 pdb structure in the RCSB PDB database. So I downloaded the PDB structure of THAP9 from the AlphaFold database.

AlphaFold Protein Structure Database



THAP9

Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli See search results





Showing all search results for THAP9

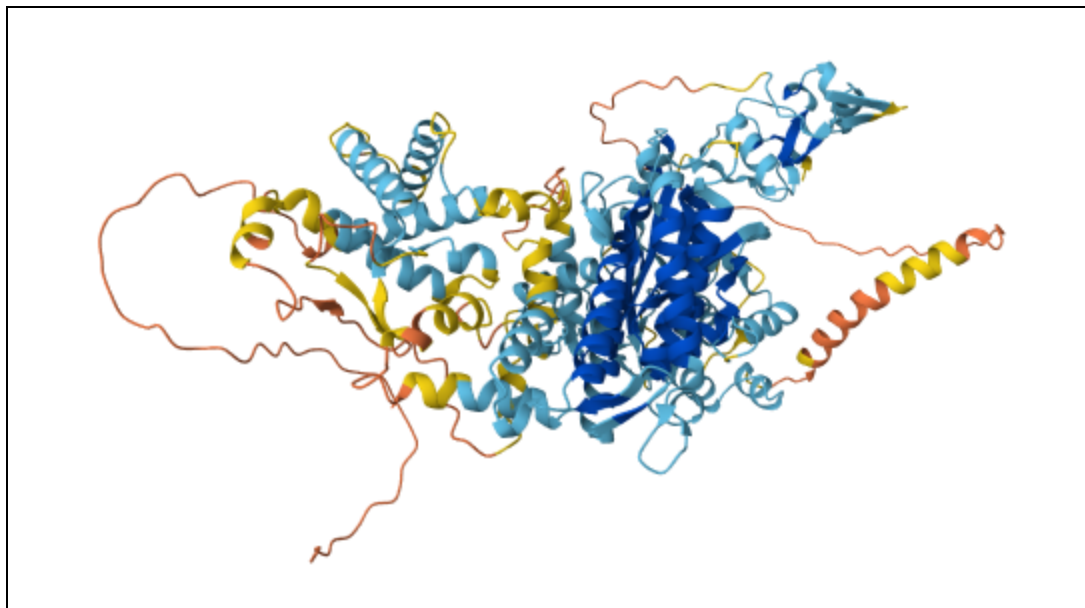
1 - 20 of 657 results

DNA transposase THAP9

AF-Q9H5L6-F1-v4 • Google DeepMind dataset •  

Tell us what you think of the new look [Share your feedback](#)

Summary and Model Confidence		Domains	Annotations	Similar Proteins
Protein	DNA transposase THAP9			Experimental structures
Gene	THAP9			Average pLDDT 
Source organism	Homo sapiens (Human) go to search 			pLDDT distribution
UniProt	Q9H5L6 go to UniProt 			<div><div></div>18.5% Very high</div>
Biological function	Active transposase that specifically recognizes the bipartite 5'-TXXGGGX(A/T)-3' consensus motif and mediates transposition. go to UniProt 			<div><div></div>44.7% High</div>
				<div><div></div>17.6% Low</div>
				<div><div></div>19.2% Very low</div>



THAP9 pdb structure

```

HEADER                                01-JUN-22
TITLE      ALPHAFOLD MONOMER V2.0 PREDICTION FOR DNA TRANSPOSASE THAP9 (Q9H5L6)
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: DNA TRANSPOSASE THAP9;
COMPND     3 CHAIN: A
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE     3 ORGANISM_TAXID: 9606
REMARK     1
REMARK     1 REFERENCE 1
REMARK     1 AUTH  JOHN JUMPER, RICHARD EVANS, ALEXANDER PRITZEL, TIM GREEN,
REMARK     1 AUTH 2 MICHAEL FIGURNOV, OLAF RONNEBERGER, KATHRYN TUNYASUVUNAKOOL,
REMARK     1 AUTH 3 RUSS BATES, AUGUSTIN ZIDEK, ANNA POTAPENKO, ALEX BRIDGLAND,
REMARK     1 AUTH 4 CLEMENS MEYER, SIMON A A KOHL, ANDREW J BALLARD,
REMARK     1 AUTH 5 ANDREW COWIE, BERNARDINO ROMERA-PAREDES, STANISLAV NIKOLOV,
REMARK     1 AUTH 6 RISHUB JAIN, JONAS ADLER, TREVOR BACK, STIG PETERSEN,
REMARK     1 AUTH 7 DAVID REIMAN, ELLEN CLANCY, MICHAL ZIELINSKI,
REMARK     1 AUTH 8 MARTIN STEINEGGER, MICHALINA PACHOLSKA, TAMAS BERGHAMMER,
REMARK     1 AUTH 9 DAVID SILVER, ORIOL VINYALS, ANDREW W SENIOR,
REMARK     1 AUTH10 KORAY KAVUKCUOGLU, PUSHMEET KOHLI, DEMIS HASSABIS
REMARK     1 TITL   HIGHLY ACCURATE PROTEIN STRUCTURE PREDICTION WITH ALPHAFOLD
REMARK     1 REF    NATURE                                V. 596   583 2021
REMARK     1 REFN   ISSN 0028-0836
REMARK     1 PMID   34265844
REMARK     1 DOI    10.1038/s41586-021-03819-2
REMARK     1

```

ATOM	1	N	MET	A	1	31.867	21.904	-41.751	1.00	51.11	N
ATOM	2	CA	MET	A	1	30.655	22.682	-41.391	1.00	51.11	C
ATOM	3	C	MET	A	1	30.962	23.625	-40.232	1.00	51.11	C
ATOM	4	CB	MET	A	1	29.480	21.766	-41.004	1.00	51.11	C
ATOM	5	O	MET	A	1	31.238	23.156	-39.134	1.00	51.11	O
ATOM	6	CG	MET	A	1	28.720	21.225	-42.217	1.00	51.11	C
ATOM	7	SD	MET	A	1	27.357	20.134	-41.741	1.00	51.11	S
ATOM	8	CE	MET	A	1	26.627	19.795	-43.369	1.00	51.11	C
ATOM	9	N	THR	A	2	30.933	24.938	-40.463	1.00	64.93	N
ATOM	10	CA	THR	A	2	31.023	25.965	-39.407	1.00	64.93	C
ATOM	11	C	THR	A	2	29.744	25.973	-38.572	1.00	64.93	C
ATOM	12	CB	THR	A	2	31.213	27.361	-40.021	1.00	64.93	C
ATOM	13	O	THR	A	2	28.651	26.085	-39.127	1.00	64.93	O
ATOM	14	CG2	THR	A	2	32.648	27.579	-40.486	1.00	64.93	C
ATOM	15	OG1	THR	A	2	30.363	27.495	-41.143	1.00	64.93	O
ATOM	16	N	ARG	A	3	29.863	25.860	-37.247	1.00	81.72	N
ATOM	17	CA	ARG	A	3	28.713	25.896	-36.331	1.00	81.72	C
ATOM	18	C	ARG	A	3	28.223	27.341	-36.173	1.00	81.72	C
ATOM	19	CB	ARG	A	3	29.097	25.276	-34.978	1.00	81.72	C
ATOM	20	O	ARG	A	3	29.037	28.251	-36.053	1.00	81.72	O
ATOM	21	CG	ARG	A	3	29.545	23.809	-35.063	1.00	81.72	C

b. Extract the following details of all C-alpha (CA) atoms of Alanine residues in chain A. (3 Points)

• Atom name, Residue name, Residue ID, Chain ID, x, y, z coordinates and save as a new alanine_info.pdb

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ awk '/^ATOM/ && $3=="CA" && $4=="ALA" && $5=="A" {print $3, $4, $5, $6, $7, $8, $9}' AF-Q9H5L6-F1-model_v4.pdb > alanine_info.pdb
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ vi alanine_info.pdb
```

```
CA ALA A 7 20.025 32.322 -30.475
CA ALA A 41 16.432 31.839 -21.892
CA ALA A 59 24.085 23.423 -30.539
CA ALA A 84 25.869 31.204 -15.763
CA ALA A 103 -6.912 33.623 -45.100
CA ALA A 120 36.349 11.021 -57.141
CA ALA A 136 58.759 8.894 -16.348
CA ALA A 140 51.818 7.502 -15.443
CA ALA A 173 21.729 -15.013 -25.800
CA ALA A 190 5.621 -15.195 -27.871
CA ALA A 210 7.163 0.561 -33.488
CA ALA A 216 12.560 -1.924 -25.535
CA ALA A 312 2.193 -11.997 -33.637
CA ALA A 318 -4.835 -9.477 -21.489
CA ALA A 326 -2.445 10.346 -8.407
CA ALA A 345 0.763 -8.631 -13.592
CA ALA A 351 7.370 -1.815 -9.768
CA ALA A 370 4.004 7.892 -1.005
CA ALA A 375 -1.760 -6.280 0.477
CA ALA A 377 1.514 -11.579 -2.974
CA ALA A 383 7.403 -3.712 -2.578
```

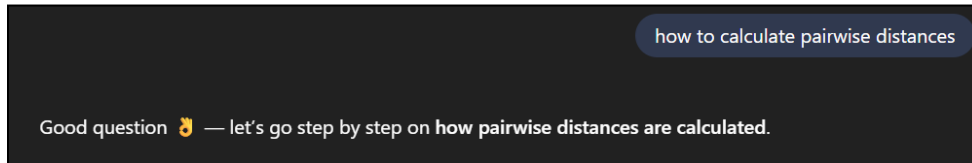
```
CA ALA A 385 11.344 -4.951 -6.142
CA ALA A 410 7.176 6.084 1.505
CA ALA A 425 -6.579 -14.274 10.684
CA ALA A 439 -8.647 -12.318 22.324
CA ALA A 449 -26.154 -14.498 18.978
CA ALA A 466 -33.386 -1.481 29.692
CA ALA A 478 -24.029 -3.651 9.948
CA ALA A 487 -15.619 -2.888 22.249
CA ALA A 489 -20.769 -3.339 24.073
CA ALA A 550 0.715 -1.872 18.267
CA ALA A 581 3.387 2.897 11.032
CA ALA A 633 -15.626 -2.424 -22.678
CA ALA A 637 -16.070 3.018 -20.060
CA ALA A 664 -19.169 -0.456 1.092
CA ALA A 670 -19.707 -6.059 9.396
CA ALA A 704 -55.378 8.701 23.909
CA ALA A 719 -41.912 10.269 9.881
CA ALA A 723 -37.364 6.482 9.017
CA ALA A 728 -32.945 0.317 4.597
```

So, in this question, I extracted the C-alpha atoms of Alanine residues in chain A. Also, print the atom name, residue name, residue id, chain id, x, y, z coordinates, and save it to a new file named alanine_info.pdb.

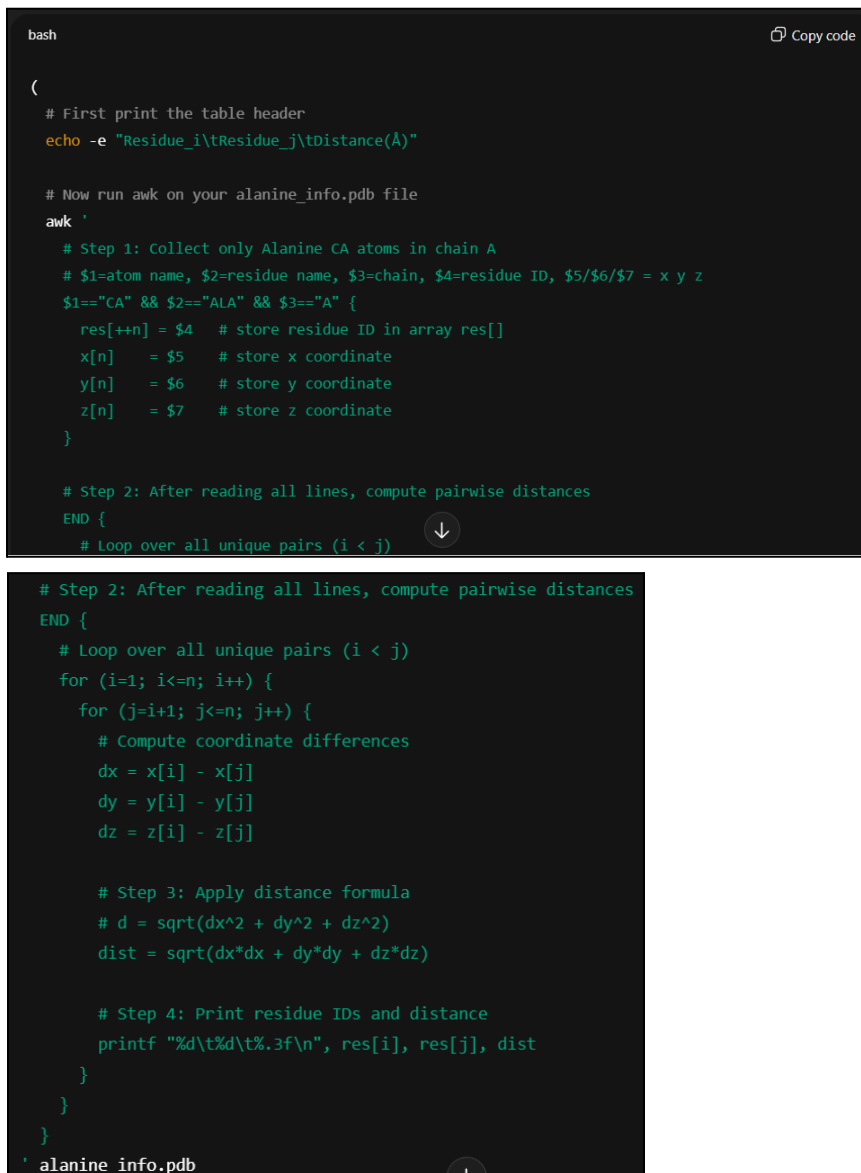
c. Using the extracted CA coordinates of Alanine, compute the pairwise distances between all alanine CA atoms in chain A. Report the results in the form of a table with the following format: (3 Points)

• Residue_i | Residue_j | Distance

Chatgpt prompt:



Chatgpt result:



Echo -e tells echo to interpret \t as a tab character. So the header columns are tab-separated. In res[++n] = \$4, ++n increments n by 1 and returns the incremented value. res[n] stores the residue ID (from field \$4) into the res array, after the first match, n==1. x[n] = \$5; y[n] = \$6; z[n] = \$7 stores the x, y, z coordinates of that CA atom into parallel arrays. dx = x[i] - x[j] (and dy, dz) compute differences in coordinates between the two residues. Here we apply the 3-D Euclidean distance formula, dist = sqrt(dx*dx + dy*dy + dz*dz).

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ vi pairwise_dist.sh
```

```
#!/bin/bash
(
  echo -e "Residue_i\tResidue_j\tDistance|"
  awk '
    $1=="CA" && $2=="ALA" && $3=="A" {
      res[++n]=$4; x[n]=$5; y[n]=$6; z[n]=$7
    }
    END {
      for(i=1;i<=n;i++){
        for(j=i+1;j<=n;j++){
          dx=x[i]-x[j]; dy=y[i]-y[j]; dz=z[i]-z[j]
          dist=sqrt(dx*dx+dy*dy+dz*dz)
          printf "%d\t%d\t%.3f\n", res[i], res[j], dist
        }
      }
    }
  ' alanine_info.pdb
)
```

```
(base) intern@rosalind:~/Nabendu/Biocomputing_Assignment/Project_assignment$ bash pairwise_dist.sh > pairwise_dist.csv
```

Residue_i	Residue_j	Distance
7	41	9.317
7	59	9.782
7	84	15.870
7	103	30.679
7	120	37.832
7	136	47.421
7	140	43.044
7	173	47.596
7	190	49.720
7	210	34.399
7	216	35.397
7	312	47.876

Github Code:

```
Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git init
Initialized empty Git repository in C:/Users/Nabendu/Downloads/Project-1-Biocomputing/.git/

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git config user.email "nabendudas"

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git config user.name "nabendudas"

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git config user.email "nabendudas001@gmail.com"
```

```
Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ pwd
/c/Users/Nabendu/Downloads/Project-1-Biocomputing

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git add /c/Users/Nabendu/Downloads/Project-1-Biocomputing
warning: in the working copy of 'README.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data/THAP1_mouse.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data/THAP1_rat.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data/THAP9_chimpanzee.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data/THAP9_human.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data/piwi_fruit_fly.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/AF-Q9H5L6-F1-model_v4.pdb', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/THAP9_coding_sequences.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/THAP9_hca.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/THAP9_homologs_5seq.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/alanine_info.pdb', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/cds.fna', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/emboss_transeq-I20250916-162342-0572-73014202-p1m.out', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/header_files.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/longest_seq.fasta', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/pairwise_dist.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'results/sequence_id5.txt', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'scripts/pairwise_dist.sh', LF will be replaced by CRLF the next time Git touches it
```

```
Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git status
On branch master

No commits yet

Changes to be committed:
  (use "git rm --cached <file>..." to unstage)
    new file:   README.txt
    new file:   data/Project_1.pdf
    new file:   data/THAP1_mouse.fasta
    new file:   data/THAP1_rat.fasta
    new file:   data/THAP9_chimpanzee.fasta
    new file:   data/THAP9_human.fasta
    new file:   data/piwi_fruit_fly.fasta
    new file:   results/AF-Q9H5L6-F1-model_v4.pdb
    new file:   results/AF-Q9H5L6-F1.png
    new file:   results/THAP9_coding_sequences.txt
    new file:   results/THAP9_hca.txt
    new file:   results/THAP9_homologs_5seq.txt
    new file:   results/alanine_info.pdb
    new file:   results/cds.fna
    new file:   results/emboss_transeq-I20250916-162342-0572-73014202-p1m.out
    new file:   results/header_files.txt
    new file:   results/longest_seq.fasta
    new file:   results/pairwise_dist.csv
    new file:   results/sequence_id5.txt
    new file:   scripts/pairwise_dist.sh
```

```

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git commit -m "/c/Users/Nabendu/Downloads/Project-1-Biocomputing"
[master (root-commit) 5f0dc3c] C:/Users/Nabendu/Downloads/Project-1-Biocomputing
20 files changed, 10161 insertions(+)
create mode 100644 README.txt
create mode 100644 data/Project_1.pdf
create mode 100644 data/THAP1_mouse.fasta
create mode 100644 data/THAP1_rat.fasta
create mode 100644 data/THAP9_chimpanzee.fasta
create mode 100644 data/THAP9_human.fasta
create mode 100644 data/piwi_fruit_fly.fasta
create mode 100644 results/AF-Q9H5L6-F1-model_v4.pdb
create mode 100644 results/AF-Q9H5L6-F1.png
create mode 100644 results/THAP9_coding_sequences.txt
create mode 100644 results/THAP9_hca.txt
create mode 100644 results/THAP9_homologs_5seq.txt
create mode 100644 results/alanine_info.pdb
create mode 100644 results/cds.fna
create mode 100644 results/emboss_transeq-I20250916-162342-0572-73014202-plm.ou
t
create mode 100644 results/header_files.txt
create mode 100644 results/longest_seq.fasta
create mode 100644 results/pairwise_dist.csv
create mode 100644 results/sequence_id5.txt
create mode 100644 scripts/pairwise_dist.sh

Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git remote add origin https://github.com/nabendudas/Project-1-Biocomputing.git


Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (master)
$ git branch -M main

```

```


Nabendu@LAPTOP-DM1JT3BQ MINGW64 ~/Downloads/Project-1-Biocomputing (main)
$ git push -u origin main
Enumerating objects: 25, done.
Counting objects: 100% (25/25), done.
Delta compression using up to 12 threads
Compressing objects: 100% (23/23), done.
Writing objects: 100% (25/25), 317.86 KiB | 3.15 MiB/s, done.
Total 25 (delta 4), reused 0 (delta 0), pack-reused 0 (from 0)
remote: Resolving deltas: 100% (4/4), done.
To https://github.com/nabendudas/Project-1-Biocomputing.git
 * [new branch]      main -> main
branch 'main' set up to track 'origin/main'.

```


Project-1-Biocomputing
Public
Pin
Watch

main
1 Branch
0 Tags

Add file
Code

 nabendudas	C:/Users/Nabendu/Downloads/Project-1-Biocomputing	5f0dc3c · 20 hours ago	1 Commit
data	C:/Users/Nabendu/Downloads/Project-1-Biocomputing	20 hours ago	
results	C:/Users/Nabendu/Downloads/Project-1-Biocomputing	20 hours ago	
scripts	C:/Users/Nabendu/Downloads/Project-1-Biocomputing	20 hours ago	
README.txt	C:/Users/Nabendu/Downloads/Project-1-Biocomputing	20 hours ago	

Github Link: <https://github.com/nabendudas/Project-1-Biocomputing/tree/main>