Neva Olliffe, PID A69026930

Find a Gene Final 12/8/23

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known.

Name: Cyclin dependent kinase 1 (CDK1)

Species: Homo Sapiens

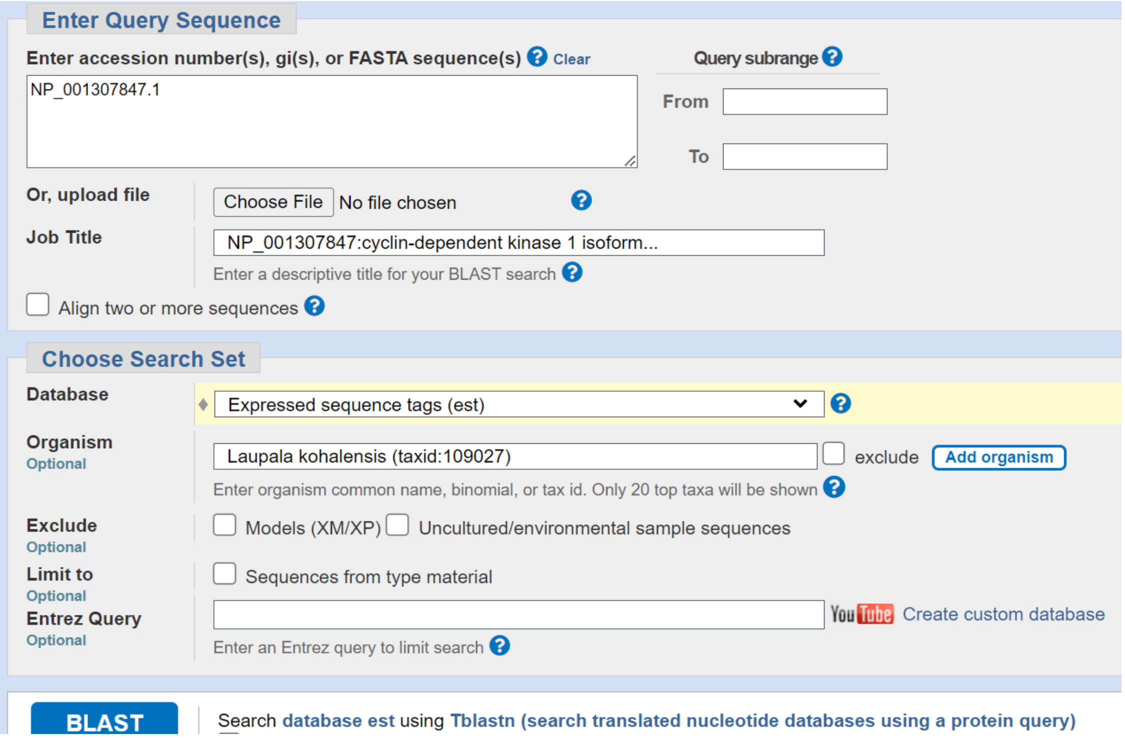
Accession number: NP_001307847.

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN (2.14.1) search against ESTs

Database: ESTs

Species: Laupala kohalensis (taxid:109027)



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

NP_001307847.1

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: NP_001307847:cyclin-dependent kinase 1 isoform...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism Optional: Laupala kohalensis (taxid:109027) □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

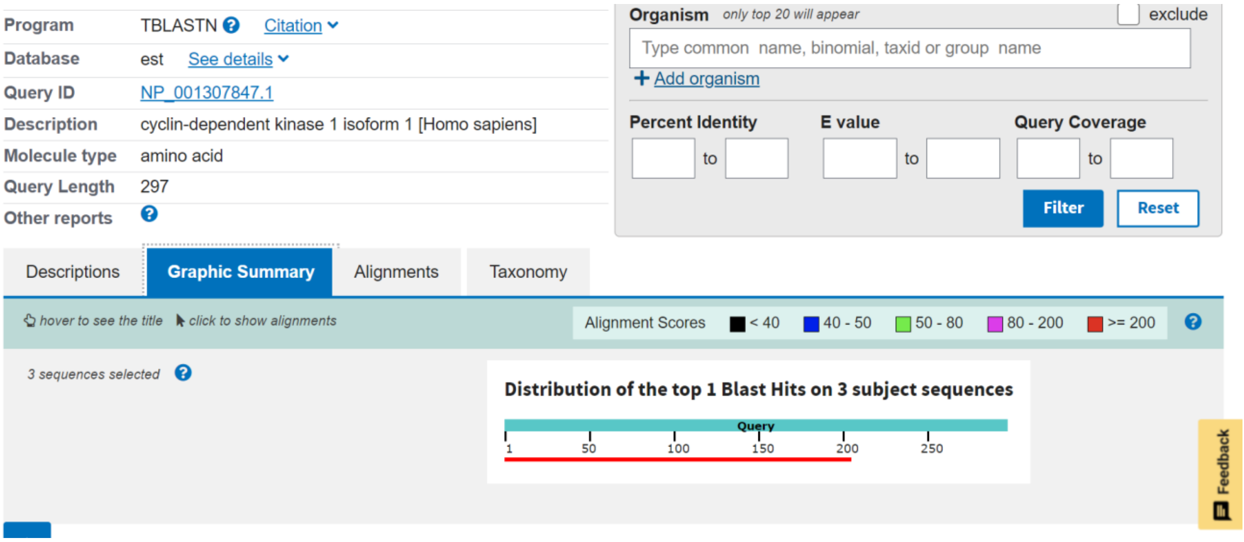
Limit to Optional: □ Sequences from type material

Entrez Query Optional: Create custom database

Enter an Entrez query to limit search

BLAST | Search database est using Tblastn (search translated nucleotide databases using a protein query)

Chosen match: Accession EH638576.1, an 846bp clone from Laupala kohalensis. Alignment details below.



Program	TBLASTN Citation
Database	est See details
Query ID	NP_001307847.1
Description	cyclin-dependent kinase 1 isoform 1 [Homo sapiens]
Molecule type	amino acid
Query Length	297
Other reports	

Organism *only top 20 will appear* □ exclude

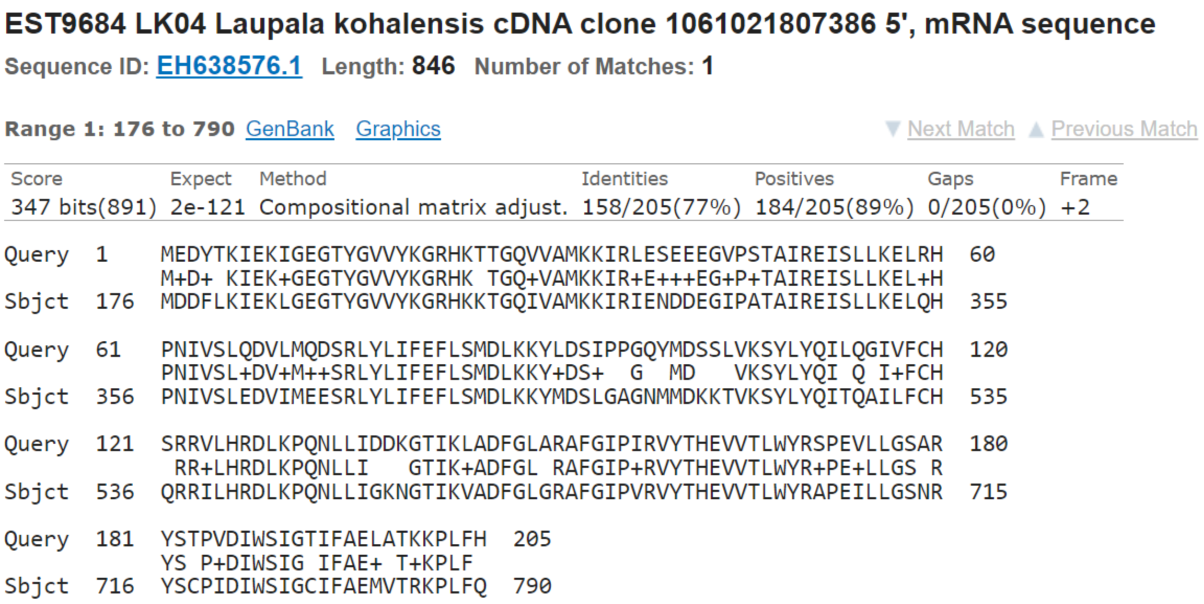
Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to E value to Query Coverage to

Filter Reset

Distribution of the top 1 Blast Hits on 3 subject sequences



EST9684 LK04 Laupala kohalensis cDNA clone 1061021807386 5', mRNA sequence

Sequence ID: EH638576.1 Length: 846 Number of Matches: 1

Range 1: 176 to 790 GenBank Graphics

```
Score          Expect   Method                      Identities      Positives       Gaps         Frame
347 bits(891)  2e-121   Compositional matrix adjust. 158/205(77%)  184/205(89%)  0/205(0%)   +2

Query  1    MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH  60
            M+D+ KIEK+GEGTYGVVYKGRHK TGQ+VAMKKIR+E+++EG+P+TAIREISLLKEL+H
Sbjct  176  MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQH  355

Query  61   PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH  120
            PNIVSL+DV+M++SRLYLIFEFLSMDLKKY+DS+  G  MD   VKSYLYQI Q I+FCH
Sbjct  356  PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCH  535

Query  121  SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR  180
             RR+LHRDLKPQNLLI   GTIK+ADFGL RAFGIP+RVYTHEVVTLWYR+PE+LLGS R
Sbjct  536  QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR  715

Query  181  YSTPVDIWSIGTIFAELATKKPLFH  205
            YS P+DIWSIG IFAE+ T+KPLF
Sbjct  716  YSCPIDIWSIGCIFAEMVTRKPLFQ  790
```

Alignment

Query: cyclin-dependent kinase 1 isoform 1 [Homo sapiens] Query ID: NP_001307847.1 Length: 297
>EST9684 LK04 Laupala kohalensis cDNA clone 1061021807386 5', mRNA sequence
Sequence ID: EH638576.1 Length: 846
Range 1: 176 to 790

Score:347 bits(891), Expect:2e-121,
Method:Compositional matrix adjust.,
Identities:158/205(77%), Positives:184/205(89%), Gaps:0/205(0%)

```
Query  1    MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH  60
            M+D+ KIEK+GEGTYGVVYKGRHK TGQ+VAMKKIR+E+++EG+P+TAIREISLLKEL+H
Sbjct  176  MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQH  355

Query  61   PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH  120
            PNIVSL+DV+M++SRLYLIFEFLSMDLKKY+DS+  G  MD   VKSYLYQI Q I+FCH
Sbjct  356  PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCH  535

Query  121  SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR  180
             RR+LHRDLKPQNLLI   GTIK+ADFGL RAFGIP+RVYTHEVVTLWYR+PE+LLGS R
Sbjct  536  QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR  715

Query  181  YSTPVDIWSIGTIFAELATKKPLFH  205
            YS P+DIWSIG IFAE+ T+KPLF
Sbjct  716  YSCPIDIWSIGCIFAEMVTRKPLFQ  790
```

[Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

> Laupala kohalensis CDK1-like protein
MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQHPNIVSLEDVIMEESRLYLIFE
FLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCHQRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPV
RVYTHEVVTLWYRAPEILLGSNRYSCPIDIWSIGCIFAEMVTRKPLFQ

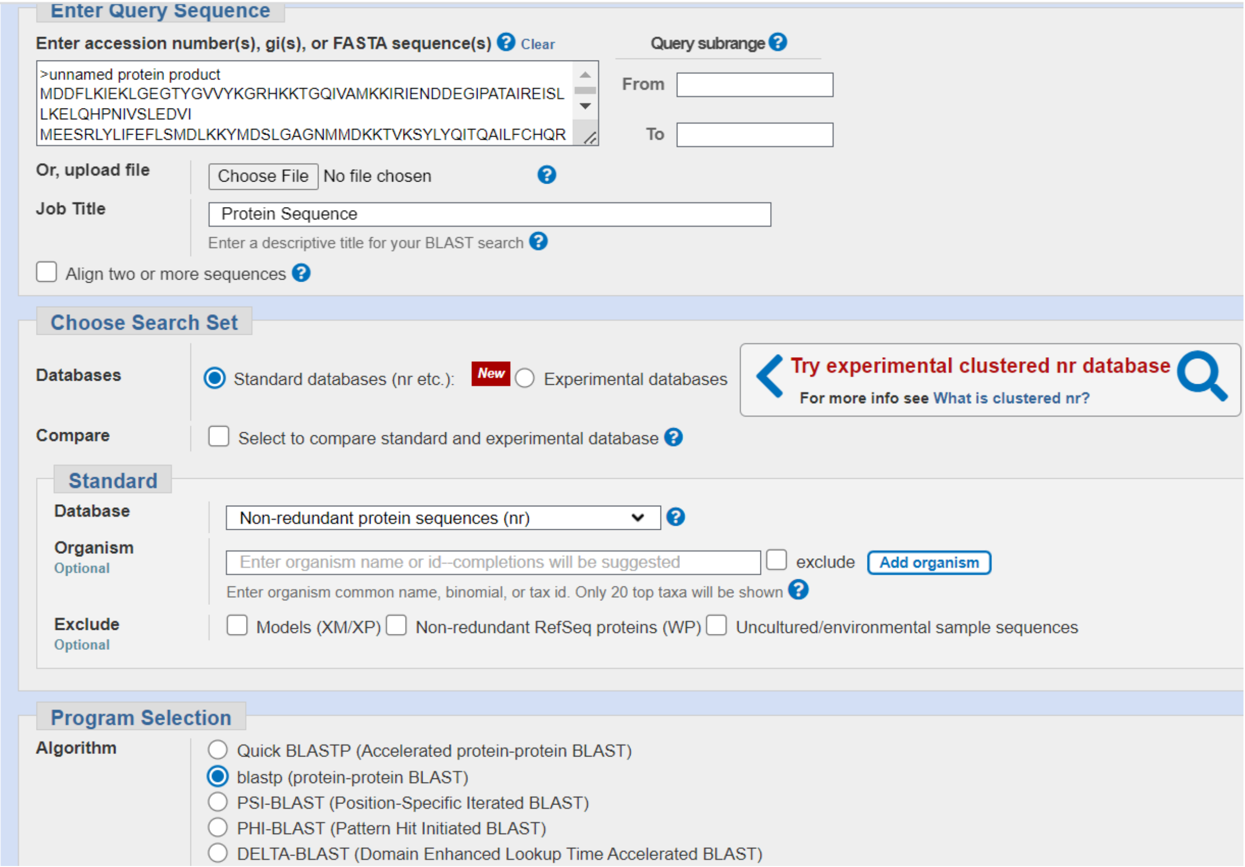
Name: Laupala kohalensis cDNA clone 1061021807386

Species: Laupala kohalensis

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Polyneoptera; Orthoptera; Ensifera; Gryllidea; Grylloidea; Gryllidae; Trigonidiinae; Laupala

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

Details: A blastp search against the protein sequence from Q3 yielded a top hit of CDK1 in Gryllus bimaculatus. Additional search results below.



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>unnamed protein product
MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISL
LKELQHPNIVSLEDVI
MEESRLYLIFEFLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCHQR
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: Protein Sequence

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

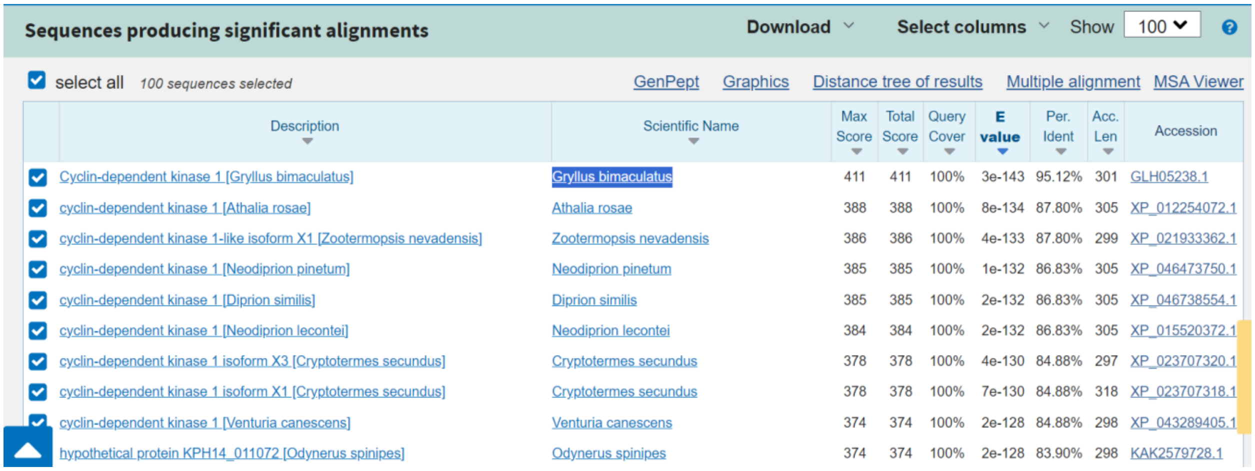
Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

The top hit was CDK1 in Gryllus bimaculatus:



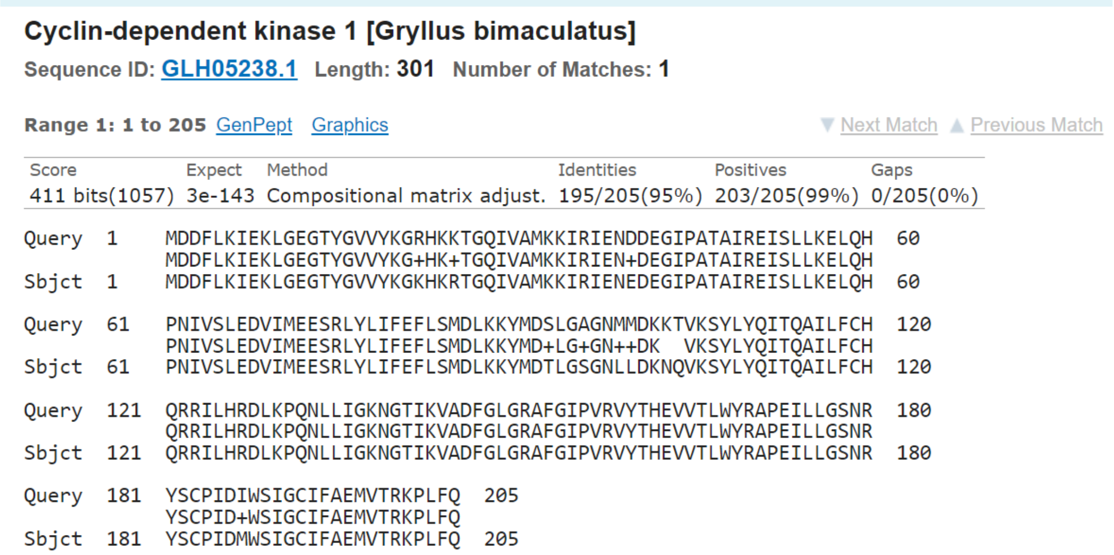
Sequences producing significant alignments

Download Select columns Show 100

☑ select all 100 sequences selected

GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Cyclin-dependent kinase 1 [Gryllus bimaculatus]	Gryllus bimaculatus	411	411	100%	3e-143	95.12%	301	GLH05238.1
☑	cyclin-dependent kinase 1 [Athalia rosae]	Athalia rosae	388	388	100%	8e-134	87.80%	305	XP_012254072.1
☑	cyclin-dependent kinase 1-like isoform X1 [Zootermopsis nevadensis]	Zootermopsis nevadensis	386	386	100%	4e-133	87.80%	299	XP_021933362.1
☑	cyclin-dependent kinase 1 [Neodiprion pinetum]	Neodiprion pinetum	385	385	100%	1e-132	86.83%	305	XP_046473750.1
☑	cyclin-dependent kinase 1 [Diprion similis]	Diprion similis	385	385	100%	2e-132	86.83%	305	XP_046738554.1
☑	cyclin-dependent kinase 1 [Neodiprion lecontei]	Neodiprion lecontei	384	384	100%	2e-132	86.83%	305	XP_015520372.1
☑	cyclin-dependent kinase 1 isoform X3 [Cryptotermes secundus]	Cryptotermes secundus	378	378	100%	4e-130	84.88%	297	XP_023707320.1
☑	cyclin-dependent kinase 1 isoform X1 [Cryptotermes secundus]	Cryptotermes secundus	378	378	100%	7e-130	84.88%	318	XP_023707318.1
☑	cyclin-dependent kinase 1 [Venturia canescens]	Venturia canescens	374	374	100%	2e-128	84.88%	298	XP_043289405.1
	hypothetical protein KPH14_011072 [Odynerus spinipes]	Odynerus spinipes	374	374	100%	2e-128	83.90%	298	KAK2579728.1



Cyclin-dependent kinase 1 [Gryllus bimaculatus]

Sequence ID: GLH05238.1 Length: 301 Number of Matches: 1

Range 1: 1 to 205 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
411 bits(1057)	3e-143	Compositional matrix adjust.	195/205(95%)	203/205(99%)	0/205(0%)

```
Query  1    MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQH  60
            MDDFLKIEKLGEGTYGVVYKG+HK+TGQIVAMKKIRIEN+DEGIPATAIREISLLKELQH
Sbjct  1    MDDFLKIEKLGEGTYGVVYKGKHKRTGQIVAMKKIRIENEDEGIPATAIREISLLKELQH  60

Query  61   PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCH  120
            PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMD+LG+GN++DK  VKSYLYQITQAILFCH
Sbjct  61   PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMDTLGSGNLLDKNQVKSYLYQITQAILFCH  120

Query  121  QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR  180
            QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR
Sbjct  121  QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR  180

Query  181  YSCPIDIWSIGCIFAEMVTRKPLFQ  205
            YSCPID+WSIGCIFAEMVTRKPLFQ
Sbjct  181  YSCPIDMWSIGCIFAEMVTRKPLFQ  205
```

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

> Re-labeled sequences for alignment

```
>Human_CDK1 ref|NP_001307847.1| cyclin-dependent kinase 1 isoform 1 [Homo sapiens]
MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRHPNIVSLQDVL
MQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCHSRRVLHRDLKPQNLLIDDKG
TIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSARYSTPVDIWSIGTIFAELATKKPLFHGDSEI
DQLFRIFRALGTPNNEVWPEVESLQDYKNTFPKWKPGSLASHVKNLDENGLDLLSKMLIYDPAKRISGKM
ALNHPYFNDLDNQIKKM


>Laupalla_CDK1
MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQHPNIVSLEDVIMEESRLYLIFEFLSMD
LKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCHQRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAP
EILLGSNRYSCPIDIWSIGCIFAEMVTRKPLFQ

>Wild_boar_CDK1 ref|NP_001152776.1| cyclin-dependent kinase 1 [Sus scrofa]
MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRHPNIVSLQDVL
MQDSRLYLIFEFLSMDLKKYLDSIPPGQFMDSSLVKSYLYQILQGIVFCHSRRVLHRDLKPQNLLIDDKG
TIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSARYSTPVDIWSIGTIFAELATKKPLFHGDSEI
DQLFRIFRALGTPNNEVWPEVESLQDYKNTFPKWKPGSLASHVKNLDENGLDLLSKMLVYDPAKRISGKM
ALNHPYFNDLDNQVKRM
```

```
>Platypus_CDK1 ref|XP_028914894.1| cyclin-dependent kinase 1 [Ornithorhynchus
anatinus]
MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRHPNIVCLQDVL
MQDARLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCHSRRVLHRDLKPQNLLIDDKG
VIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSARYSTPVDIWSIGTIFAELATKKPLFHGDSEI
DQLFRIFRALGTPNNEVWPEVESLQDYKNTFPKWKPGSLASHVKNLDENGIDLLSKMLVYDPAKRISGKM
ALNHPYFNDLDKFNLPSSQIKKF

>Drosophila_CDK1 ref|XP_041450630.1| cyclin-dependent kinase 1 isoform X2 [Drosophila
obscura]
MEDFEKIEKIGEGTYGVVYKGRNRLTGQIVAMKKIRLESDDEGVPSTAIREISLLKELKHSNIVCLEDVL
MEENRIYLIFEFLSMDLKKYMDSLPPEKLMDSKLVRSYLFQITSAILFCHRRRVLHRDLKPQNLLIDKNG
IIKVADFGLGRSFGIPVRIYTHEIVTLWYRAPEVLLGSPRYSCPVDIWSIGCIFAEMATRKPLFQEFSKL
QLKTFGQALLRFPIIKILFLAGQQIN

>Zebrafish_CDK1 ref|NP_997729.1| cyclin-dependent kinase 1 [Danio rerio]
MDDYLKIEKIGEGTYGVVYKGRNKTTGQVVAMKKIRLESEEEGVPSTAVREISLLKELQHPNVVRLLDVL
MQESKLYLVFEFLSMDLKKYLDSIPSGEFMDPMLVKSYLYQILEGILFCHCRRVLHRDLKPQNLLIDNKG
VIKLADFGLARAFGVPVRVYTHEVVTLWYRAPEVLLGASRYSTPVDLWSIGTIFAELATKKPLFHGDSEI
DQLFRIFRTLGTPNNEVWPDVESLPDYKNTFPKWKSGNLANTVKNLDKNGIDLLMKMLIYDPPKRISARQ
AMTHPYFDDLDKSSLPASNLKI

>Stegodyphus_CDK1 ref|XP_035229147.1| cyclin-dependent kinase 1-like [Stegodyphus
dumicola]
MEDYVKVEKIGEGTYGVVYKGKHKKTGRIVALKKIRIENEDEGVPSTALREISTLKELNHPNVVALLDVL
MQESRLYLVFEFLSMDLKKYLDSIPSGQFMDKALVKSYMYQLLEGILFCHRRRYLHRDLKPQNLLIDEKG
VIKIADFGLARAFGIPVRVYTHEVVTLWYRAPEVLLGSPRYSTPVDIWSAGCIFAEMANKTPLFRGDSEI
DQLFRIFRTMGTPTEDMWPGVTQLPDFKTSFPNWKSKSLSVLTTRLGSAGQALLEEMLVYNPGERISAKE
ALQHEYFDDFDKSSLPFYSPETVF
```

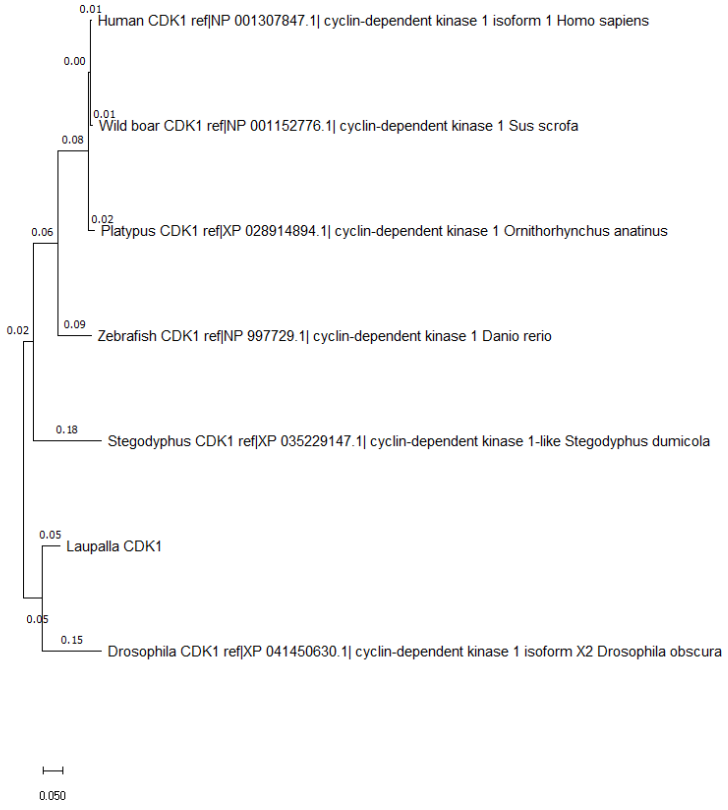
Alignment
Obtained using MUSCLE from ebi

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Stegodyphus_CDK1      MEDYVKVEKIGEGTYGVVYKGKHKKTGRIVALKKIRIENEDEGVPSTALREISTLKELNH
Zebrafish_CDK1        MDDYLKIEKIGEGTYGVVYKGRNKTTGQVVAMKKIRLESEEEGVPSTAVREISLLKELQH
Platypus_CDK1         MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH
Human_CDK1            MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH
Wild_boar_CDK1        MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH
Drosophila_CDK1       MEDFEKIEKIGEGTYGVVYKGRNRLTGQIVAMKKIRLESDDEGVPSTAIREISLLKELKH
Laupalla_CDK1         MDDFLKIEKLGEGTYGVVYKGRHKKTGQIVAMKKIRIENDDEGIPATAIREISLLKELQH
                      *:*: *:**:***********.:. **.:**:****:*.::**:*:**:**** ****.*

Stegodyphus_CDK1      PNVVALLDVLMQESRLYLVFEFLSMDLKKYLDSIPSGQFMDKALVKSYMYQLLEGILFCH
Zebrafish_CDK1        PNVVRLLDVLMQESKLYLVFEFLSMDLKKYLDSIPSGEFMDPMLVKSYLYQILEGILFCH
Platypus_CDK1         PNIVCLQDVLMQDARLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH
Human_CDK1            PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH
Wild_boar_CDK1        PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQFMDSSLVKSYLYQILQGIVFCH
Drosophila_CDK1       SNIVCLEDVLMEENRIYLIFEFLSMDLKKYMDSLPPEKLMDSKLVRSYLFQITSAILFCH
Laupalla_CDK1         PNIVSLEDVIMEESRLYLIFEFLSMDLKKYMDSLGAGNMMDKKTVKSYLYQITQAILFCH
                      .*:* * **:*:: .:**:***********:**: . : **   *.**::*: ..*:***

Stegodyphus_CDK1      RRRYLHRDLKPQNLLIDEKGVIKIADFGLARAFGIPVRVYTHEVVTLWYRAPEVLLGSPR
Zebrafish_CDK1        CRRVLHRDLKPQNLLIDNKGVIKLADFGLARAFGVPVRVYTHEVVTLWYRAPEVLLGASR
Platypus_CDK1         SRRVLHRDLKPQNLLIDDKGVIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR
Human_CDK1            SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR
Wild_boar_CDK1        SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR
Drosophila_CDK1       RRRVLHRDLKPQNLLIDKNGIIKVADFGLGRSFGIPVRIYTHEIVTLWYRAPEVLLGSPR
Laupalla_CDK1         QRRILHRDLKPQNLLIGKNGTIKVADFGLGRAFGIPVRVYTHEVVTLWYRAPEILLGSNR
                       ** ************..:* **:*****.*:**:*:*:****:******:**:***: *

Stegodyphus_CDK1      YSTPVDIWSAGCIFAEMANKTPLFRGDSEIDQLFRIFRTMGTPTEDMWPGVTQLPDFKTS
Zebrafish_CDK1        YSTPVDLWSIGTIFAELATKKPLFHGDSEIDQLFRIFRTLGTPNNEVWPDVESLPDYKNT
```

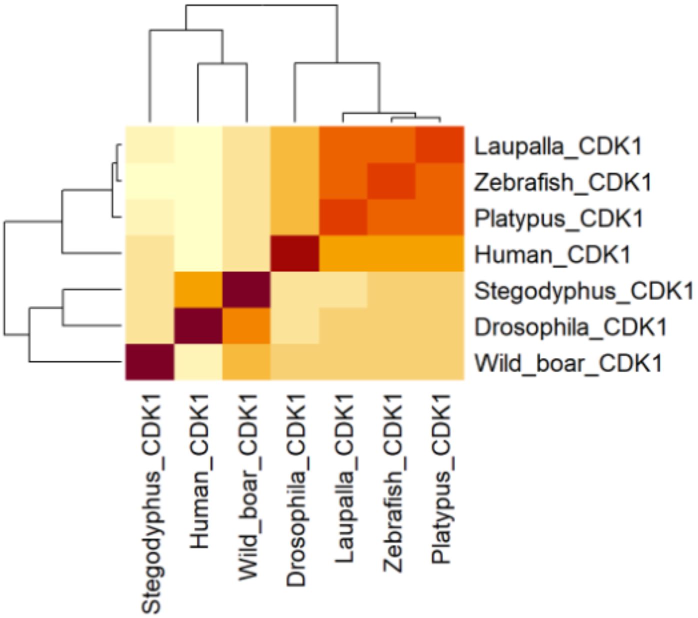
```
Platypus_CDK1         YSTPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESLQDYKNT
Human_CDK1            YSTPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESLQDYKNT
Wild_boar_CDK1        YSTPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESLQDYKNT
Drosophila_CDK1       YSCPVDIWSIGCIFAEMATRKPLFQEFSKL------------------------------
Laupalla_CDK1         YSCPIDIWSIGCIFAEMVTRKPLFQ-----------------------------------
                      ** *:*:** * ****:....***.

Stegodyphus_CDK1      FPNWKSKSLSVLTTRLGSAGQALLEEMLVYNPGERISAKEALQHEYFDDFDKSSLPFYSP
Zebrafish_CDK1        FPKWKSGNLANTVKNLDKNGIDLLMKMLIYDPPKRISARQAMTHPYFDDLDKSSLPASNL
Platypus_CDK1         FPKWKPGSLASHVKNLDENGIDLLSKMLVYDPAKRISGKMALNHPYFNDLDKFNLPSSQI
Human_CDK1            FPKWKPGSLASHVKNLDENGLDLLSKMLIYDPAKRISGKMALNHPYFNDLD------NQI
Wild_boar_CDK1        FPKWKPGSLASHVKNLDENGLDLLSKMLVYDPAKRISGKMALNHPYFNDLD------NQV
Drosophila_CDK1       -----------QLKTFGQALLRFPIIKILFLAGQQIN-----------------------
Laupalla_CDK1         ------------------------------------------------------------


Stegodyphus_CDK1      ETVF
Zebrafish_CDK1        KI--
Platypus_CDK1         KKF-
Human_CDK1            KKM-
Wild_boar_CDK1        KRM-
Drosophila_CDK1       ----
Laupalla_CDK1         ----
```

[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Imported previous sequences into MEGA, aligned using the MUSCLE algorithm, and created a neighbor joining tree.


0.01
Human CDK1 ref|NP 001307847.1| cyclin-dependent kinase 1 isoform 1 Homo sapiens
0.00
0.01
Wild boar CDK1 ref|NP 001152776.1| cyclin-dependent kinase 1 Sus scrofa
0.08
0.02
Platypus CDK1 ref|XP 028914894.1| cyclin-dependent kinase 1 Ornithorhynchus anatinus
0.06
0.09
Zebrafish CDK1 ref|NP 997729.1| cyclin-dependent kinase 1 Danio rerio
0.02
0.18
Stegodyphus CDK1 ref|XP 035229147.1| cyclin-dependent kinase 1-like Stegodyphus dumicola
0.05
Laupalla CDK1
0.05
0.15
Drosophila CDK1 ref|XP 041450630.1| cyclin-dependent kinase 1 isoform X2 Drosophila obscura
0.050

[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.



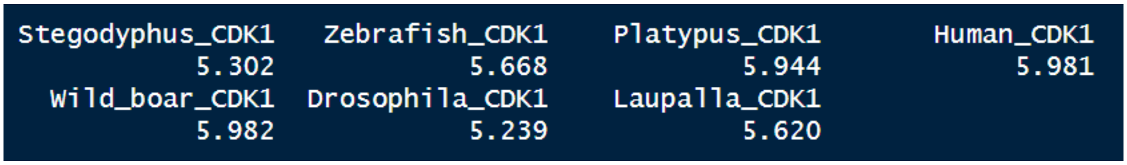
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.
List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimental Technique), resolution (resolution), and source organism (source).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task.
Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above. Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could choose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

Consensus sequence generated with bio3d:

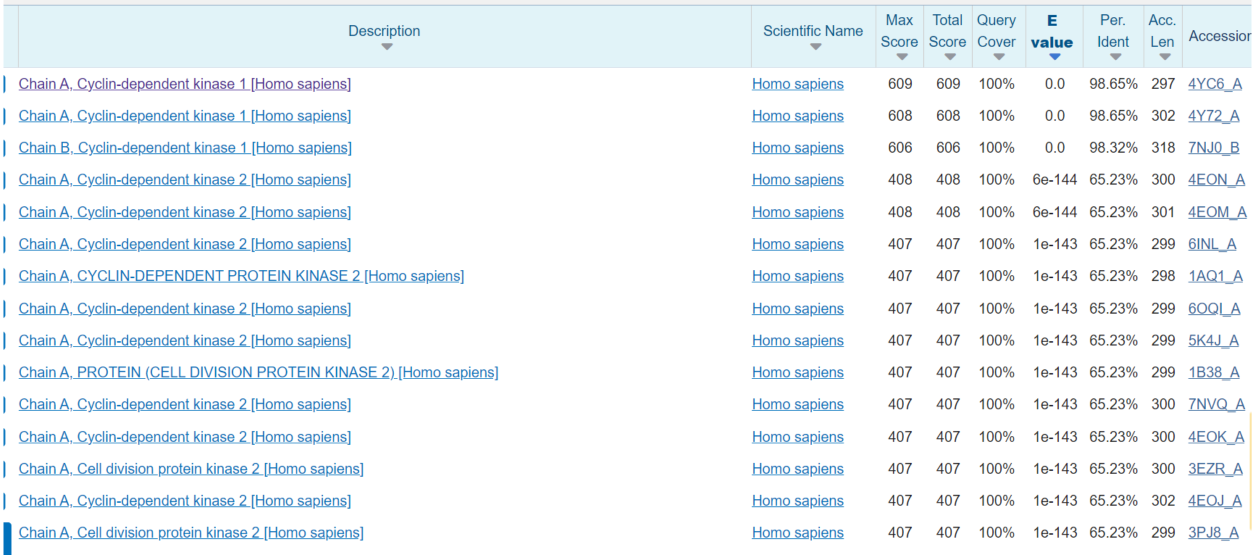
```
  [1] "M" "E" "D" "Y" "-" "K" "I" "E" "K" "I" "G" "E" "G" "T" "Y" "G" "V" "V"
 [19] "Y" "K" "G" "R" "H" "K" "-" "T" "G" "Q" "-" "V" "A" "M" "K" "K" "I" "R"
 [37] "L" "E" "S" "E" "-" "E" "G" "V" "P" "S" "T" "A" "I" "R" "E" "I" "S" "L"
 [55] "L" "K" "E" "L" "-" "H" "P" "N" "I" "V" "-" "L" "-" "D" "V" "L" "M" "Q"
 [73] "-" "S" "R" "L" "Y" "L" "I" "F" "E" "F" "L" "S" "M" "D" "L" "K" "K" "Y"
 [91] "L" "D" "S" "I" "P" "-" "G" "-" "-" "M" "D" "-" "-" "L" "V" "K" "S" "Y"
[109] "L" "Y" "Q" "I" "L" "-" "G" "I" "-" "F" "C" "H" "-" "R" "R" "V" "L" "H"
[127] "R" "D" "L" "K" "P" "Q" "N" "L" "L" "I" "D" "-" "K" "G" "-" "I" "K" "-"
[145] "A" "D" "F" "G" "L" "A" "R" "A" "F" "G" "I" "P" "-" "R" "V" "Y" "T" "H"
[163] "E" "V" "V" "T" "L" "W" "Y" "R" "-" "P" "E" "V" "L" "L" "G" "S" "-" "R"
[181] "Y" "S" "T" "P" "V" "D" "I" "W" "S" "I" "G" "-" "I" "F" "A" "E" "-" "A"
[199] "T" "K" "K" "P" "L" "F" "-" "G" "D" "S" "E" "I" "D" "Q" "L" "F" "R" "I"
[217] "F" "R" "-" "-" "G" "T" "P" "-" "-" "-" "-" "W" "P" "-" "V" "-" "-" "L"
[235] "-" "D" "-" "K" "-" "-" "F" "P" "-" "W" "K" "-" "-" "-" "L" "-" "-" "-"
[253] "-" "K" "-" "L" "-" "-" "-" "G" "-" "-" "L" "L" "-" "-" "M" "L" "-" "Y"
[271] "-" "P" "-" "-" "R" "I" "S" "-" "-" "-" "A" "-" "-" "H" "-" "Y" "F" "-"
[289] "D" "-" "D" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
```

There are a lot of gaps in the last ~100 residues, so I will move forward with a single sequence instead of the consensus sequence. The Wild boar CDK1 sequence has the highest identity to other sequences, so I will proceed with the wild board sequence.



```
Stegodyphus_CDK1    Zebrafish_CDK1     Platypus_CDK1        Human_CDK1
           5.302             5.668             5.944             5.981
  Wild_boar_CDK1  Drosophila_CDK1     Laupalla_CDK1
           5.982             5.239             5.620
```

Wild boar CDK1 was entered into protein BLAST against the PDB database. Results summary:



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accessior
Chain A, Cyclin-dependent kinase 1 [Homo sapiens]	Homo sapiens	609	609	100%	0.0	98.65%	297	4YC6_A
Chain A, Cyclin-dependent kinase 1 [Homo sapiens]	Homo sapiens	608	608	100%	0.0	98.65%	302	4Y72_A
Chain B, Cyclin-dependent kinase 1 [Homo sapiens]	Homo sapiens	606	606	100%	0.0	98.32%	318	7NJ0_B
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	408	408	100%	6e-144	65.23%	300	4EON_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	408	408	100%	6e-144	65.23%	301	4EOM_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	299	6INL_A
Chain A, CYCLIN-DEPENDENT PROTEIN KINASE 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	298	1AQ1_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	299	6OQI_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	299	5K4J_A
Chain A, PROTEIN (CELL DIVISION PROTEIN KINASE 2) [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	299	1B38_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	300	7NVQ_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	300	4EOK_A
Chain A, Cell division protein kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	300	3EZR_A
Chain A, Cyclin-dependent kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	302	4EOJ_A
Chain A, Cell division protein kinase 2 [Homo sapiens]	Homo sapiens	407	407	100%	1e-143	65.23%	299	3PJ8_A

Top 3 results:

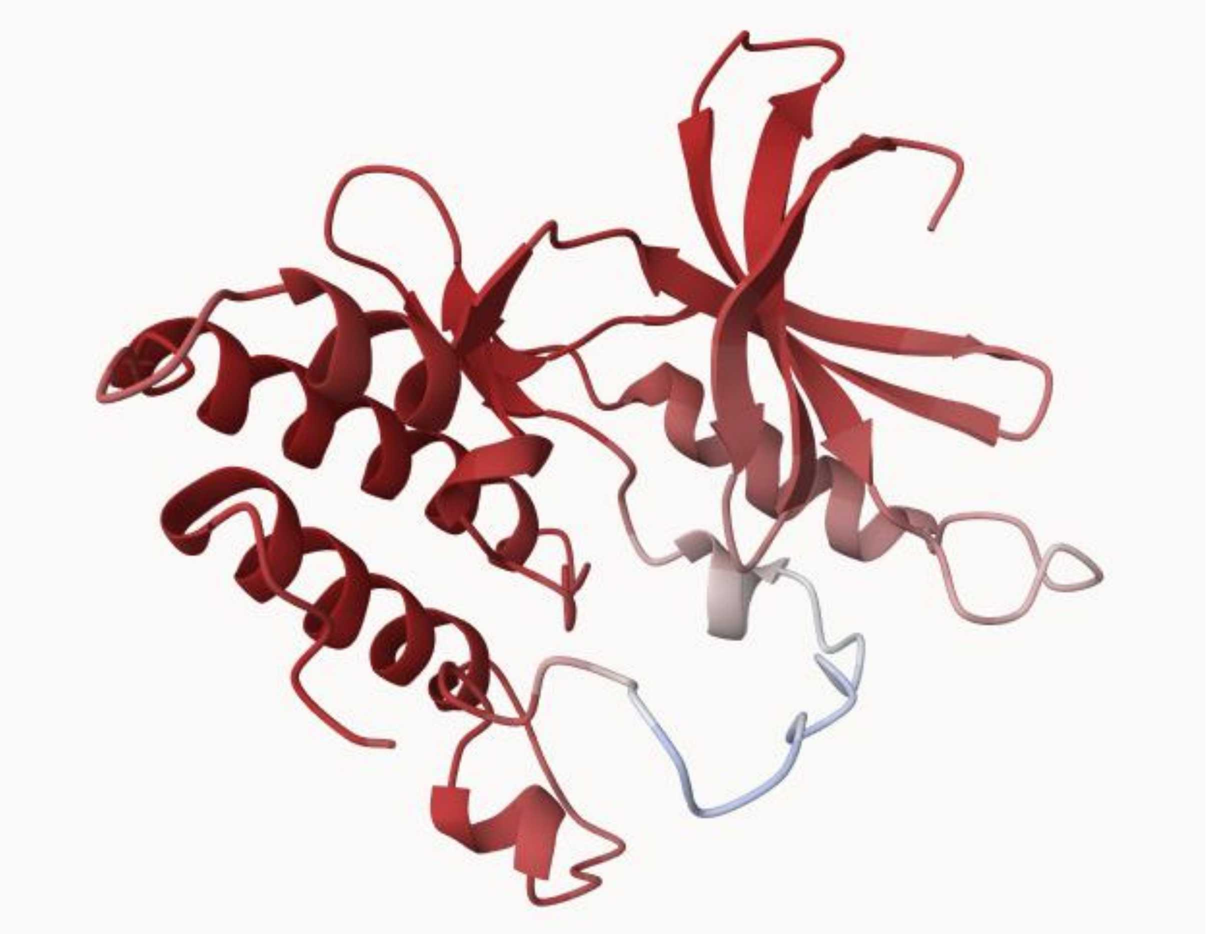
ID	Technique	Resolution	Source	Evalue	Identity
4YC6	X-ray diffraction	2.6Å	Homo sapiens	0.0	98.65
4Y72	X-ray diffraction	2.3Å	Homo sapiens	0.0	98.65
7NJ0	X-ray diffraction	3.6Å	Homo sapiens	0.0	98.32

[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

> Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a "too many amino acids" (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

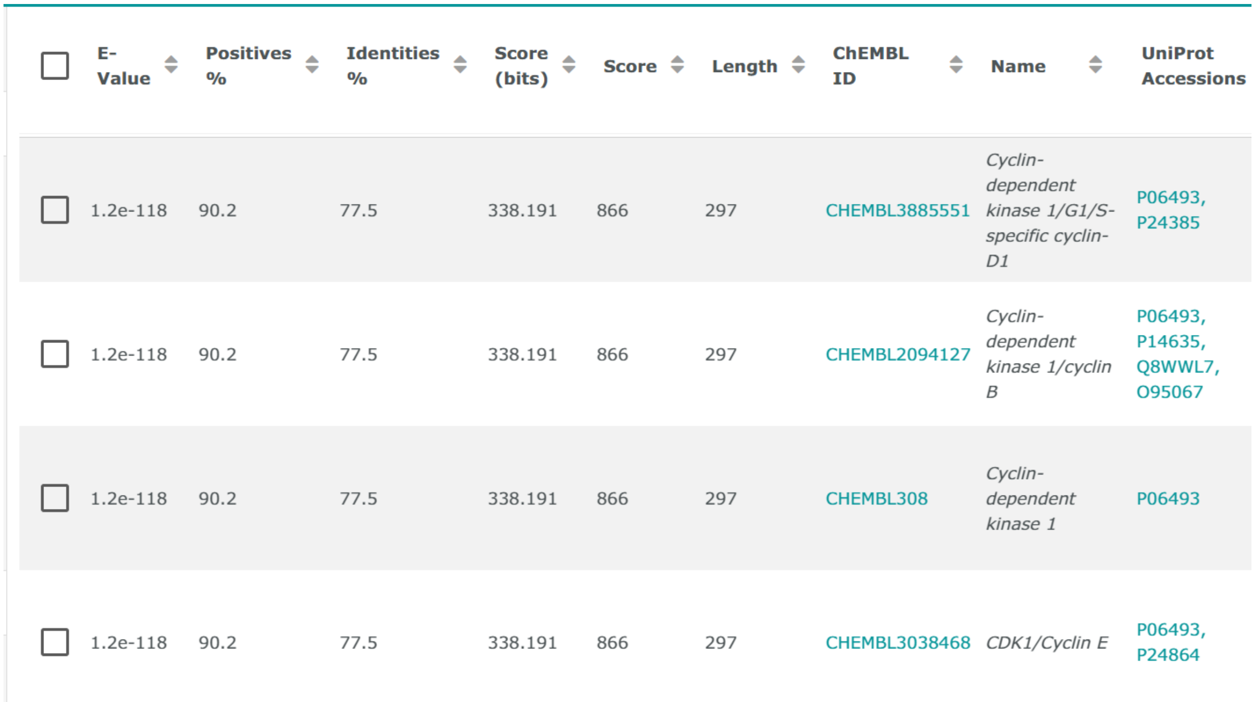
Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the Mol* viewer online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight conserved residues that are likely to be functional as spacefill and the protein as cartoon colored by local alpha fold pLDDT quality score. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).

Laupala kohalensis CDK1-like structure visualized in Mol*, colored by pLDDT quality score.



[Q10] Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list "non available as of [date]".

Top ChEMBL search results:



□	E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions
□	1.2e-118	90.2	77.5	338.191	866	297	CHEMBL3885551	*Cyclin-dependent kinase 1/G1/S-specific cyclin-D1*	P06493, P24385
□	1.2e-118	90.2	77.5	338.191	866	297	CHEMBL2094127	*Cyclin-dependent kinase 1/cyclin B*	P06493, P14635, Q8WWL7, O95067
□	1.2e-118	90.2	77.5	338.191	866	297	CHEMBL308	*Cyclin-dependent kinase 1*	P06493
□	1.2e-118	90.2	77.5	338.191	866	297	CHEMBL3038468	*CDK1/Cyclin E*	P06493, P24864

Note that the *Mus musculus* single protein listing was 10th in the search result. This is because there were a lot of results for "protein-protein interaction" and "protein complex" for CDK1 and various cyclin binding partners. I chose to only report on the "single molecule" results for CDK1 alone, since I did not look into what cyclins in *Laupala kohalensis* are highly conserved with human cyclins.

The ChEMBL search identified CDK1 in both *Homo sapiens* (CHEMBL307) and *Mus musculus* (CHEMBL4084). In mice, ChEMBL identified one binding assay and ligand efficiency data for 6 ligands. For human CDK1, ChEMBL identified 357 binding assays, 7 functional assays, and 1 toxicity assay. There is ligand efficiency data for 1,189 molecules. Given that *Laupala kohalensis* shares 77% identity of human CDK1, it is highly likely that one of the many existing binding or functional assays would be useful for measuring inhibition of *Laupala* CDK1. Additionally, some of the nearly 1200 assayed ligands are likely to bind and potentially inhibit *Laupala* CDK1 as well.