Bioinformatics, 2021, 03–22
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: n/a
Manuscript Category

Sequence analysis

CDK5 in C. Elegans: Analysis of the Cyclin Dependent Kinase 5 in Caenorhabditis Elegans

Jacob Casey[1,*]

[1]School of Computing, Newcastle University, Address 1 Science Square, Newcastle upon Tyne NE4 5TG

*To whom correspondence should be addressed.

Abstract

Motivation: The UniProt Knowledgebase is a far from complete source of genomic annotation and thus requires algorithmic computation to verify and predict further annotations. This work analyses the CDK5 in C. elegans to further strengthen the UniProt Swiss-Prot annotation and find gaps in the knowledgebase.

Results: Confirmed the 6-exon sequence structure of the protein that exists. CDK5 is a homogenous to many eukaryotic cells and confirmed that the CDK5 transcript belongs to the kinase-like family. Confirmed the ATP binding site at 33 and binding region between 10-18. Discovered potential ATP binding site at position 31 and modeled the structure of a sequence in Darwinula stevensoni (currently unannotated) that likely contains the CDK5 protein.

Availability: Daily 9am – 5:30pm.

Contact: j.casey@newcastle.ac.uk

1 Introduction

1.1 UniProt Knowledgebase and the Relationship Between Swiss-Prot and TrEMBL

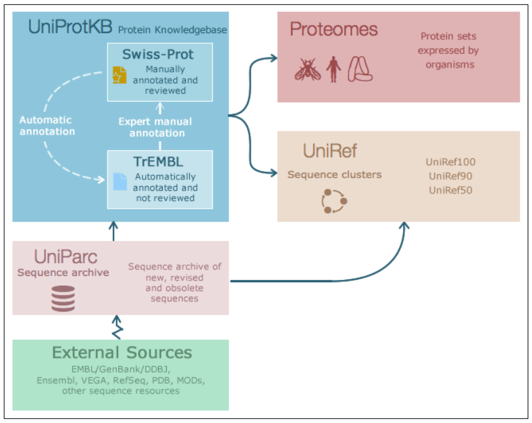

UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ, Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

Fig. 1. UniProt Knowledgebase Structure. This shows the breakdown and relations of protein resources. Source: https://www.uniprot.org/help/about

UniProt is a comprehensive, high-quality and freely accessible resource of protein information (UniProt, 2021). This includes protein sequences and functions, and protein annotations (Flegel, 2021). The UniProt protein knowledgebase (Ebi, 2021) is split up into two sets - TrEMBL and Swiss-Prot. With the TrEMBL set, the protein sequences are automatically annotated and are not reviewed and are thus not considered as the final and correct protein sequence. TrEMBL entries are made from UniParc, UniProt's archive dataset that contains most of all available protein sequences by providing a unique identifier to that sequence. Furthermore, UniParc, is derived from many external sources such as Ensembl, WormBase and EMBL-Bank. Finally, the other set that makes up Uniprot is Swiss-Prot. The entries are manually annotated and are reviewed by biocurators (SCQ, 2021) to achieve protein sequences with low redundancy and as a result of this, Swiss-Prot tends to be a more reliable source of a proteins true sequence.

Due to this discrepancy between Swiss-Prot and TrEMBL databases, one cannot assume that the sequences are accurate or complete, even if they have been annotated by an expert. It is therefore the job of the bioinformatician to check, where necessary, the accuracy of the proteins, to analyze them to find any potential gaps in genomic data in the current knowledge base.

1.2 Protein Sequences

Five proteins were assigned for potential analysis. CDK5_CAEEL https://www.uniprot.org/uniprot/G5ECH7 belongs to Caenorhabditis elegans (C. elegans) and is involved in several motor neurons and promotes the polarized trafficking of synaptic vesicles and dense core vesicles (DCV). The UniProt annotation is reviewed, 5-star with experimental evidence at protein level. It features in 18 publications. Also, there exists a homologue in homo sapiens with 228 publications.

128UP_DROME https://www.uniprot.org/uniprot/P32234 belongs to Drosophila melanogaster (Fruit fly) and interacts with the gene Deformed (Dfd) to be activated in maxillary segment cells. The UniProt status is reviewed, with a 3-star annotation score with experimental evidence at transcript level. It features in 29 publications.

BACC_BACIU https://www.uniprot.org/uniprot/Q8KWT4 belongs to the Gram-positive Bacillus subtilis bacterium and is responsible partly for the biosynthesis of bacilysin in the bacABCDEFG operon. The UniProt status is as reviewed, an annotation score of 2-star with experimental evidence at protein level. It features in one publication.

A0A4S5AQX5_ECOLI https://www.uniprot.org/uniprot/A0A4S5AQX5 for Escherichia coli (E. coli) strain K12 is a left amino acid, and its function is related to N-acetyltransferase activity. The UniProt summary is unreviewed, 1-star annotation score with the protein predicted. This sequence is derived from an EMBL/GenBank/DDBJ whole genome shotgun (WGS) entry. It features in one publication.

A0A5S9YCB0_ARATH https://www.uniprot.org/uniprot/A0A5S9YCB0 is an uncharacterised protein for Arabidopsis thaliana (Mouse-ear cress) and its function is unknown. It has the UniProt annotation status of 1-star with protein predicted. It features in one publication.

Clearly the two of most interest because of more features in publications and being more richly annotated on UniProt were CDK5_CAEEL and 128UP_DROME. Due to the better UniProt annotation of CDK5_CAEEL this was picked. Specifically, being 5-star reviewed with experimental evidence at the protein level compared to 3-star at experimental evidence at transcript level of 128UP_DROME as this offers more chance to compare and verify the results of UniProt.

1.3 CDK5 in C. Elegans UniProt Annotations that Currently Exist

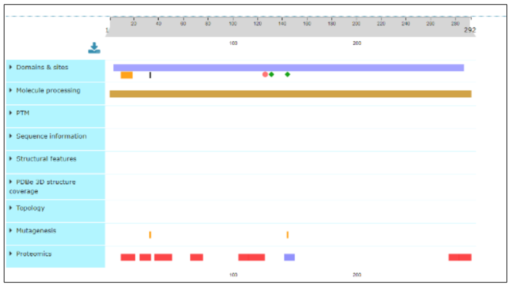


Fig. 1. UniProt feature view of CDK5 C. elegans

It is of length 292 aa. UniProt has annotated binding sites: 33 for ATP, 126 for proton acceptor, 131 for magnesium and 144 also for magnesium. There is also a nucleotide binding region between 10-18 for ATP. The molecular function is identified as of the kinase type. Mutagenesis sites are identified at position 33 as K → T and 144 as D → N.

2 Methods

The methods were started with the assumption that the UniProt annotation was not known. This way, this work could independently predict gene annotations to enable the comparison with these results and those from the UniProt team.

2.1 Gene Prediction

Used the Ensembl gene summary: https://metazoa.ensembl.org/Caenorhabditis_elegans/Gene/Summary?db=core;g=WB Gene00000407;r=III:13464687-13466390;t=T27E9.3.1 reference to https://wormbase.org/species/c_elegans/gene/WBGene00000407 to get the gene transcripts. The CDK5 gene has 1 transcript (ID: T27E9.3.1) with 6 exons with 1,105 base pairs and 292 amino acids.

Table 1. Ensembl Export

Gene to export	Output	Strand	5' Flanking sequence	3' Flanking sequence
WBGene 0000040 7 (cdk-5)	FASTA sequence	Feature Strand	1000	1000

Parameters for exporting the gene. Saved as .txt file.

This exported .txt was inputted to FGENESH (http://linux1.softberry.com/organism) specific gene-finding parameter set to C. elegans and advanced options left as default: only print mRNA sequences as true.

BLAST on UniProt (https://www.uniprot.org/blast/) was used to compare the predicted proteins from FGENESH.

Table 2. BLAST parameters for identifying FGENESH similarity

Database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB	10	Auto	None	Yes	250

Next geneid 1.2 hosted on Web Server 2005 at genome.crg.es (https://genome.crg.es/geneid.html) was used for an alternative gene prediction result. The same sequence from Table 1. Was given to the following geneid parameter setup.

Table 3. Geneid parameters

Organism	Prediction mode	DNA strands	Output format
C. elegans	Normal	Forward and Reverse	GFF

Input was FASTA format derived from Table 1.

2.2 Homologues

NCBI-BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to search protein sequence databases to identify where the sequence appears in other species and with what similarity. The FASTA format was taken of the CDK5 protein from UniProt and entered into the BLASTp algorithm.

Table 4. BLASTp general parameters for identifying homologs

Database	Max Target Sequences	Expect Threshold	Word Size	Max Matches
Non-redundant	100	0.05	6	0

BLASTp Tree View was used to visualize the most similar sequences returned by BLASTp within a distance of 0.15 using the fast tree method with Kimura distance.

2.3 Guide Trees and Phylogenetic Trees

EBI Clustal-Omega https://www.ebi.ac.uk/Tools/msa/clustalo/ was used with the sequences with the top 9 most similar sequences from BLASTp. Output format was set as ClustalW with character counts. 9 was chosen to match the tree view results within 0.15 distance.

2.4 Gene Function

BLASTp multiple sequence alignment with same parameters as 2.2 Homologues section. The protein classification graphical summary was used with concise mode.

NsitePred (http://biomine.cs.vcu.edu/servers/NsitePred/) was used to search multiple sequences for their nucleotide binding sites to then compare results to the BLASTp graphical representation and UniProt annotation. The top 5 similar sequences (Maximum number NsitePred allowed) from the BLASTp result were given as FASTA format. This is the only parameter the NsitePred takes.

MOTIF Search https://www.genome.jp/tools/motif/ with the 100 most similar sequences identified by BLASTp with the following parameters.

Table 5. MOTIF Search input parameters

Database	Cut-off score	Skip entries with SKIP-FLAG	Skip unspecific profiles
Pfam	1.0	True	True

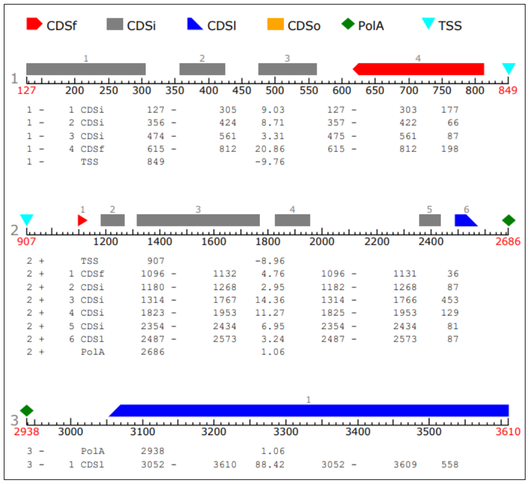
2.5 Protein Structure Prediction

PSIPRED GenTHREADER (http://bioinf.cs.ucl.ac.uk/ psipred/) with the CDK5_CAEEL sequence data in FASTA format with default parameter settings. Visualized and aligned with PyMOL2 (https://pymol.org/2/). ChimeraX https://www.rbvi.ucsf.edu/chimerax/ was used with the seqview tool to inspect the number of helix and strand sections in the predicted structures. ChimeraX atom view to show nucleotide structure.

Modeler v10.1 on Windows 10 was used to align and model the Darwinula stevensoni (seed shrimp) homologous sequence. ChainA from the https://www.rcsb.org/structure/3o0g entry (PDB ID: 3o0g) for CDK5 C. elegans real protein model. Visualized, aligned and chain A selected with PyMOL2.

3 Results

3.1 Gene Prediction



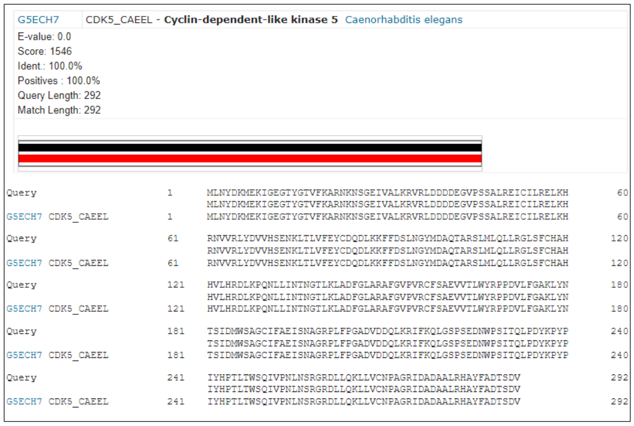
```
1 -    1 CDSi      127 -       305    9.03       127 -       303    177
1 -    2 CDSi      356 -       424    8.71       357 -       422     66
1 -    3 CDSi      474 -       561    3.31       475 -       561     87
1 -    4 CDSf      615 -       812   20.86       615 -       812    198
1 -      TSS       849               -9.76
```

```
2 +      TSS       907               -8.96
2 +    1 CDSf     1096 -      1132    4.76      1096 -      1131     36
2 +    2 CDSi     1180 -      1268    2.95      1182 -      1268     87
2 +    3 CDSi     1314 -      1767   14.36      1314 -      1766    453
2 +    4 CDSi     1823 -      1953   11.27      1825 -      1953    129
2 +    5 CDSi     2354 -      2434    6.95      2354 -      2434     81
2 +    6 CDSl     2487 -      2573    3.24      2487 -      2573     87
2 +      PolA     2686                1.06
```

```
3 -      PolA     2938                1.06
3 -    1 CDSl     3052 -      3610   88.42      3052 -      3609    558
```

Fig. 2. FGENESH predicted genes.

The results of FGENESH predicted in total 3 genes with 11 exons. The first gene has 3 internal exons: 127-305, 356-424, 474-561. It also has a exon with a start codon at the end: 615-812. The second gene has 4 internal exons: 1180-1268, 1314-1767, 1823-1953, 2354-2434. It also has the starting and ending exons:1096-1132 and 2487-2573 respectively and in the correct order. The third gene has only a last coding segment: 3052-3610.



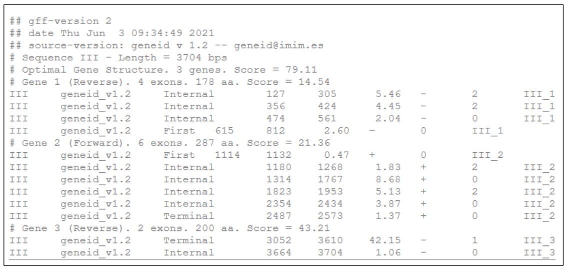
G5ECH7 CDK5_CAEEL - **Cyclin-dependent-like kinase 5** Caenorhabditis elegans

E-value: 0.0
Score: 1546
Ident.: 100.0%
Positives : 100.0%
Query Length: 292
Match Length: 292

```
Query                1     MLNYDKMEKIGEGTYGTVFKARNKNSGEIVALKRVRLDDDDEGVPSSALREICILRELKH     60
                           MLNYDKMEKIGEGTYGTVFKARNKNSGEIVALKRVRLDDDDEGVPSSALREICILRELKH
G5ECH7 CDK5_CAEEL     1     MLNYDKMEKIGEGTYGTVFKARNKNSGEIVALKRVRLDDDDEGVPSSALREICILRELKH     60

Query                61    RNVVRLYDVVHSENKLTLVFEYCDQDLKKFFDSLNGYMDAQTARSLMLQLLRGLSFCHAH    120
                           RNVVRLYDVVHSENKLTLVFEYCDQDLKKFFDSLNGYMDAQTARSLMLQLLRGLSFCHAH
G5ECH7 CDK5_CAEEL     61    RNVVRLYDVVHSENKLTLVFEYCDQDLKKFFDSLNGYMDAQTARSLMLQLLRGLSFCHAH    120

Query                121   HVLHRDLKPQNLLINTNGTLKLADFGLARAFGVPVRCFSAEVVTLWYRPPDVLFGAKLYN    180
                           HVLHRDLKPQNLLINTNGTLKLADFGLARAFGVPVRCFSAEVVTLWYRPPDVLFGAKLYN
G5ECH7 CDK5_CAEEL     121   HVLHRDLKPQNLLINTNGTLKLADFGLARAFGVPVRCFSAEVVTLWYRPPDVLFGAKLYN    180

Query                181   TSIDMWSAGCIFAEISNAGRPLFPGADVDDQLKRIFKQLGSPSEDNWPSITQLPDYKPYP    240
                           TSIDMWSAGCIFAEISNAGRPLFPGADVDDQLKRIFKQLGSPSEDNWPSITQLPDYKPYP
G5ECH7 CDK5_CAEEL     181   TSIDMWSAGCIFAEISNAGRPLFPGADVDDQLKRIFKQLGSPSEDNWPSITQLPDYKPYP    240

Query                241   IYHPTLTWSQIVPNLNSRGRDLLQKLLVCNPAGRIDADAALRHAYFADTSDV             292
                           IYHPTLTWSQIVPNLNSRGRDLLQKLLVCNPAGRIDADAALRHAYFADTSDV
G5ECH7 CDK5_CAEEL     241   IYHPTLTWSQIVPNLNSRGRDLLQKLLVCNPAGRIDADAALRHAYFADTSDV             292
```

Fig. 3. UniProt BLAST result with predicted protein sequence from FGENESH.

The predicted protein (number two in Fig. 2) by FGENESH aligns with an identicalness score of 100% to the Swiss-Prot annotated CDK5 protein to that in UniProt. The length is also equal in both instances at 292 giving a score of 1546.



```
## gff-version 2
## date Thu Jun  3 09:34:49 2021
## source-version: geneid v 1.2 -- geneid@imim.es
# Sequence III - Length = 3704 bps
# Optimal Gene Structure. 3 genes. Score = 79.11
# Gene 1 (Reverse). 4 exons. 178 aa. Score = 14.54
III     geneid_v1.2     Internal        127     305      5.46   -       2       III_1
III     geneid_v1.2     Internal        356     424      4.45   -       2       III_1
III     geneid_v1.2     Internal        474     561      2.04   -       0       III_1
III     geneid_v1.2     First   615     812      2.60   -       0       III_1
# Gene 2 (Forward). 6 exons. 287 aa. Score = 21.36
III     geneid_v1.2     First   1114    1132     0.47   +       0       III_2
III     geneid_v1.2     Internal        1180    1268     1.83   +       2       III_2
III     geneid_v1.2     Internal        1314    1767     8.68   +       0       III_2
III     geneid_v1.2     Internal        1823    1953     5.13   +       2       III_2
III     geneid_v1.2     Internal        2354    2434     3.87   +       0       III_2
III     geneid_v1.2     Terminal        2487    2573     1.37   +       0       III_2
# Gene 3 (Reverse). 2 exons. 200 aa. Score = 43.21
III     geneid_v1.2     Terminal        3052    3610    42.15   -       1       III_3
III     geneid_v1.2     Internal        3664    3704     1.06   -       0       III_3
```

Fig. 4. Geneid GFF output.

Geneid found the similar type of gene structure with 3 genes predicted with this time 12, as opposed to 11, exons predicted. The start and stop positions have some overlap, particularly in gene 1 and gene 3, with gene 2 only have 1 identical exon as seen in exon 5. The length of gene 1 is the same as in FGENESH, however gene 2 is no longer of the correct aa of 292 but instead at 287 aa.

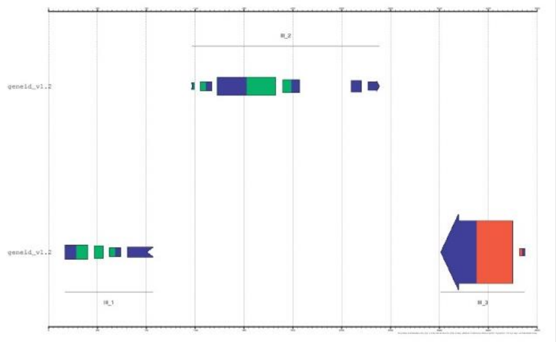
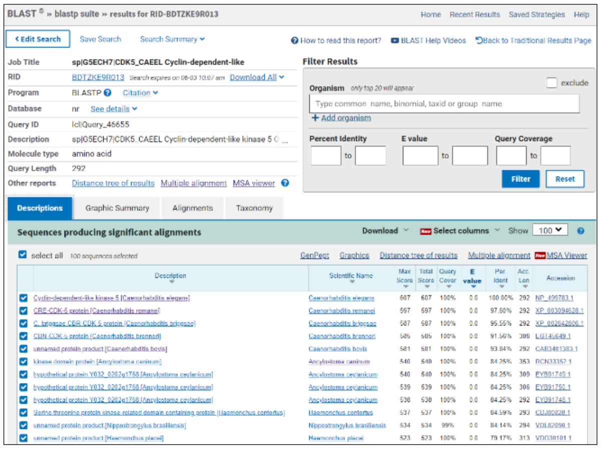


Fig. 5. Visualization output of predicted gene region from Geneid.

As can be seen in in Fig. 5 the three genes. The second gene closely corresponds to that of the FGENESH except that it appears the first section has been cut off in geneid. The starting gene on FGENESH ends at 1132 whereas the geneid gene starts at 1132, suggesting that geneid missed the start position (TATA box).

3.2 Homologues



BLAST ® » blastp suite » results for RID-BDTZKE9R013

Job Title	sp|G5ECH7|CDK5_CAEEL Cyclin-dependent-like
RID	BDTZKE9R013 Search expires on 06-03 10:07 am Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_46655
Description	sp|G5ECH7|CDK5_CAEEL Cyclin-dependent-like kinase 5 O ...
Molecule type	amino acid
Query Length	292
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

☐ exclude

Organism only top 20 will appear

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ☐ to ☐ E value ☐ to ☐ Query Coverage ☐ to ☐

Sequences producing significant alignments

☑ select all 100 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Cyclin-dependent-like kinase 5 [Caenorhabditis elegans]	Caenorhabditis elegans	607	607	100%	0.0	100.00%	292	NP_499783.1
☑	CRE-CDK-5 protein [Caenorhabditis remanei]	Caenorhabditis remanei	597	597	100%	0.0	97.60%	292	XP_003094628.1
☑	C. briggsae CBR-CDK-5 protein [Caenorhabditis briggsae]	Caenorhabditis briggsae	587	587	100%	0.0	95.55%	292	XP_002642806.1
☑	CBN-CDK-5 protein [Caenorhabditis brenneri]	Caenorhabditis brenneri	585	585	100%	0.0	91.56%	308	EGT46649.1
☑	unnamed protein product [Caenorhabditis bovis]	Caenorhabditis bovis	581	581	100%	0.0	93.84%	292	CAB3401383.1
☑	kinase domain protein [Ancylostoma caninum]	Ancylostoma caninum	540	540	100%	0.0	84.25%	353	RCN33352.1
☑	hypothetical protein Y032_0202g1768 [Ancylostoma ceylanicum]	Ancylostoma ceylanicum	540	540	100%	0.0	84.25%	309	EYB91749.1
☑	hypothetical protein Y032_0202g1768 [Ancylostoma ceylanicum]	Ancylostoma ceylanicum	539	539	100%	0.0	84.25%	306	EYB91750.1
☑	hypothetical protein Y032_0202g1768 [Ancylostoma ceylanicum]	Ancylostoma ceylanicum	538	538	100%	0.0	84.25%	292	EYB91748.1
☑	Serine threonine protein kinase-related domain containing protein [Haemonchus contortus]	Haemonchus contortus	537	537	100%	0.0	84.59%	293	CDJ80828.1
☑	unnamed protein product [Nippostrongylus brasiliensis]	Nippostrongylus brasiliensis	534	534	99%	0.0	84.14%	294	VDL82090.1
☑	unnamed protein product [Haemonchus placei]	Haemonchus placei	523	523	100%	0.0	79.17%	313	VDO30101.1

Fig. 6. NCBI-BLASTp results page.

BLASTp successfully found the corresponding entry in UniProt with 100% identity. There were many near hits with the top 100 entries being between 97.5%-75.52% similar. 94/100 top entries belonged to the animal family Ecdysozoa with the closest matches being of the Caenorhabditis genus. Furthermore, it is to point out that a match in Caenorhabditis bovis, also of length 292 is found, which is currently in the knowledge base as an "unnamed protein product" it can be said with confidence that this protein is in fact CDK5.

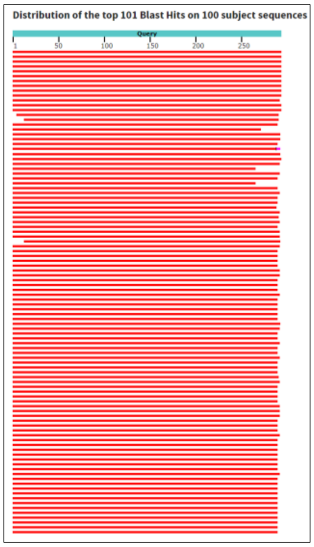


Fig. 7. Visualization of distribution of top 101 BLAST hits.

To analyze Fig. 7, it can be identified that practically every sequence here is very similar. They all share the same length except for a few, very minor inconsistencies but most line up almost perfectly.

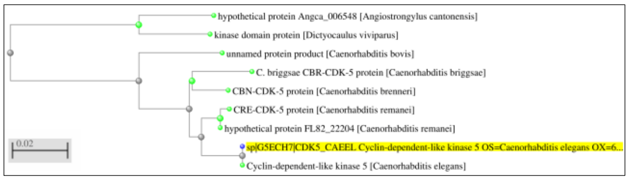

hypothetical protein Angca_006548 [Angiostrongylus cantonensis]
kinase domain protein [Dictyocaulus viviparus]
unnamed protein product [Caenorhabditis bovis]
C. briggsae CBR-CDK-5 protein [Caenorhabditis briggsae]
CBN-CDK-5 protein [Caenorhabditis brenneri]
CRE-CDK-5 protein [Caenorhabditis remanei]
hypothetical protein FL82_22204 [Caenorhabditis remanei]
sp|G5ECH7|CDK5_CAEEL Cyclin-dependent-like kinase 5 OS=Caenorhabditis elegans OX=6...
Cyclin-dependent-like kinase 5 [Caenorhabditis elegans]
0.02

Fig. 8. BLASTp Tree View.

This results in 9 sequences within 0.15 distance of the CDK5_CAEEL highlighted in yellow. The most closely related are from Caenorhabditis genus again with also results from Dictyocaulus viviparus and Angiostrongylul cantonensis. All 9 of which are from nematodes family.

3.3 Guide Trees and Phylogenetic Trees

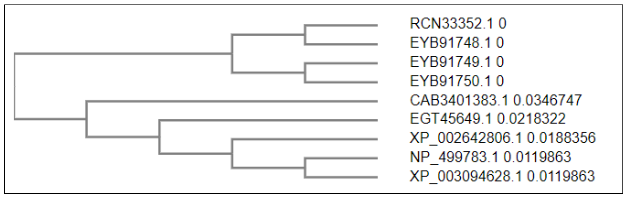


Fig. 9. Clustal Omega guide tree.

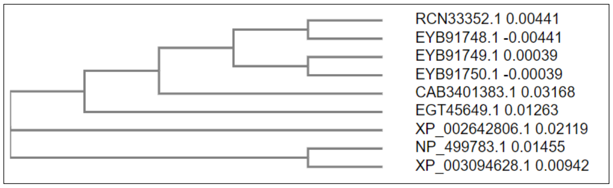


Fig. 10. Clustal Omega phylogenetic tree.

EMBL-EBI clustal omega returns the above guide tree Fig 6. and phylogenetic tree in Fig. 7. Both are different to the BLASTp Tree View results.

3.4 Gene Function

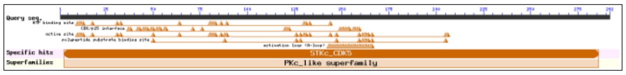
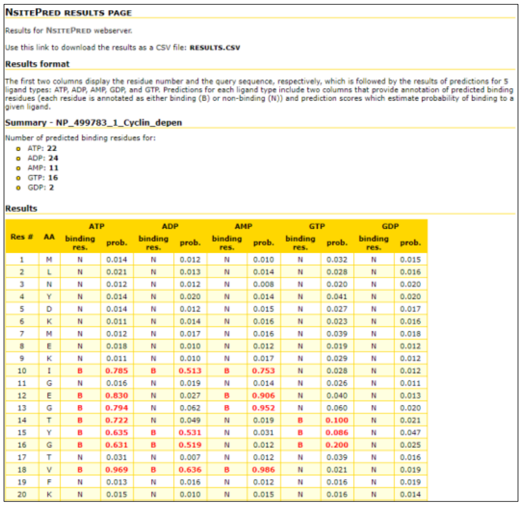


Fig. 11. BLASTp protein classification graphical summary

Gene function belongs to protein kinase superfamily. Also, can be seen that there are many binding sites throughout. There is a clear activation loop (A-loop) at 143-168.

NsitePred gives each sequence amino acids's (aa) a probability of each individual nucleotide being a binding site. On all of the sequences there is a clear binding region between 10-18 as can be seen in Fig 12.



NsitePred results page

Results for NsitePred webserver.

Use this link to download the results as a CSV file: **RESULTS.CSV**

Results format

The first two columns display the residue number and the query sequence, respectively, which is followed by the results of predictions for 5 ligand types: ATP, ADP, AMP, GDP, and GTP. Predictions for each ligand type include two columns that provide annotation of predicted binding residues (each residue is annotated as either binding (B) or non-binding (N)) and prediction scores which estimate probability of binding to a given ligand.

Summary - NP_499783_1_Cyclin_depen

Number of predicted binding residues for:

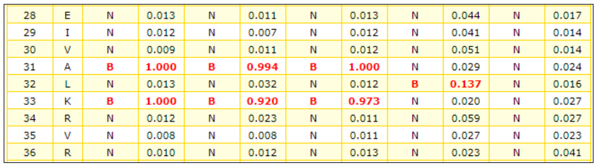
- ATP: **22**
- ADP: **24**
- AMP: **11**
- GTP: **16**
- GDP: **2**

Results

		ATP		ADP		AMP		GTP		GDP	
Res #	AA	binding res.	prob.	binding res.	prob.	binding res.	prob.	binding res.	prob.	binding res.	prob.
1	M	N	0.014	N	0.012	N	0.010	N	0.032	N	0.015
2	L	N	0.021	N	0.013	N	0.014	N	0.028	N	0.016
3	N	N	0.012	N	0.012	N	0.008	N	0.020	N	0.020
4	Y	N	0.014	N	0.020	N	0.014	N	0.041	N	0.020
5	D	N	0.014	N	0.012	N	0.015	N	0.027	N	0.017
6	K	N	0.011	N	0.014	N	0.016	N	0.023	N	0.016
7	M	N	0.012	N	0.017	N	0.016	N	0.039	N	0.018
8	E	N	0.018	N	0.010	N	0.012	N	0.019	N	0.012
9	K	N	0.011	N	0.010	N	0.017	N	0.029	N	0.012
10	I	B	0.785	B	0.513	B	0.753	N	0.028	N	0.012
11	G	N	0.016	N	0.019	N	0.014	N	0.026	N	0.011
12	E	B	0.830	N	0.027	B	0.906	N	0.040	N	0.013
13	G	B	0.794	N	0.062	B	0.952	N	0.060	N	0.020
14	T	B	0.722	N	0.049	N	0.019	B	0.100	N	0.021
15	Y	B	0.635	B	0.531	N	0.031	B	0.086	N	0.047
16	G	B	0.631	B	0.519	N	0.012	B	0.200	N	0.025
17	T	N	0.031	N	0.007	N	0.012	N	0.039	N	0.016
18	V	B	0.969	B	0.636	B	0.986	N	0.021	N	0.019
19	F	N	0.013	N	0.016	N	0.012	N	0.016	N	0.019
20	K	N	0.015	N	0.010	N	0.015	N	0.016	N	0.014

Fig. 12. Results page of NsitePred. Red indications are binding sites predicted. Only slows first 1-20 aa in the first sequence which goes on for the full 292 aa.

Also, in all the similar sequences identified by BLASTp there is an ATP binding site predicted at location 31 and 33 with a probability of 1 along with AMP binding site predicted at position 31 for all the sequences checked.



28	E	N	0.013	N	0.011	N	0.013	N	0.044	N	0.017
29	I	N	0.012	N	0.007	N	0.012	N	0.041	N	0.014
30	V	N	0.009	N	0.011	N	0.012	N	0.051	N	0.014
31	A	B	1.000	B	0.994	B	1.000	N	0.029	N	0.024
32	L	N	0.013	N	0.032	N	0.012	B	0.137	N	0.016
33	K	B	1.000	B	0.920	B	0.973	N	0.020	N	0.027
34	R	N	0.012	N	0.023	N	0.011	N	0.059	N	0.027
35	V	N	0.008	N	0.008	N	0.011	N	0.027	N	0.023
36	R	N	0.010	N	0.012	N	0.013	N	0.023	N	0.041

Fig. 13. Sequence section 28-36 snippet of full results.

Next was done was the MOTIF Search. There were many similar results, and some interesting ones. All the 100 sequences were identified to belong to the Pkinase family.

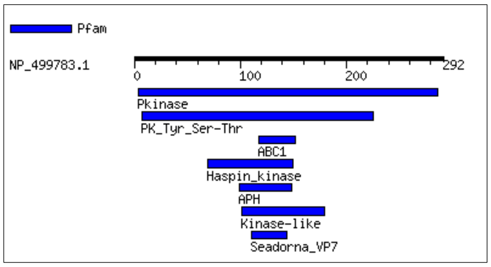


Fig. 14. CDK5 C. elegans result searching Pfam for motifs. 7 motifs identified.

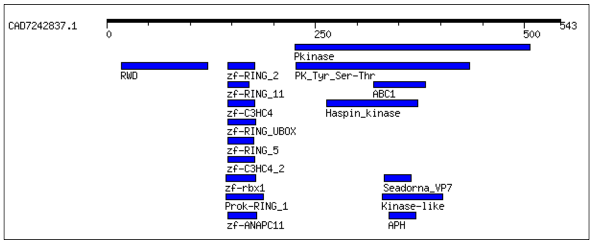
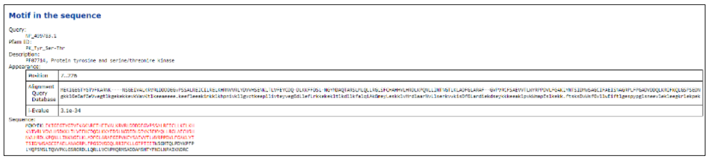


Fig. 15. Darwinula stevensoni result searching Pfam for motifs. 17 motifs found.

An interesting result from Darwinula stevensoni (seed shrimp) gives the Protein kinase family along with some other zf-RING motifs and a RWD motif. This gene has a score of 78% identities with the CDK5 from the BLASTp search.

The two closest motifs e values: Pkinase(3.5e-70) and PK_Tyr_Ser-Thr (3.1e-34) for the C. elegan gene suggest that highly that the gene is a protein tyrosine and serine/threonine kinase gene and transfers a phosphate group from ATP to a protein.



Motif in the sequence

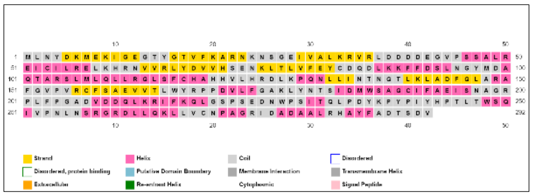
Query:
NP_499783.1
Pfam ID:
PK_Tyr_Ser-Thr
Description:
PF07714, Protein tyrosine and serine/threonine kinase
Appearance:

Position	7..276
Alignment Query Database	[illegible]
I-Evalue	3.1e-34

Sequence:
[illegible]

Fig 16. Specific result for PK_Tyr_Ser-Thr motif in CDK5 gene.

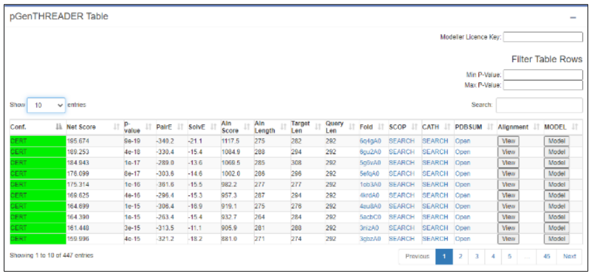
3.5 Protein Structure Prediction



```
        10        20        30        40        50
  1 MLNYDKMEKIGEGTYGTVFKARNKNSGEIVALKRVRLDDDDEGVPSSALR  50
 51 EICILRELKHRNVVRLYDVVHSENKLTLVFEYCDQDLKKFFDSLNGYMDA 100
101 QTARSLMLQLLRGLSFCHAHHVLHRDLKPQNLLINTNGTLKLADFGLARA 150
151 FGVPVRCFSAEVVTLWYRPPDVLFGAKLYNTSIDMWSAGCIFAEISNAGR 200
201 PLFPGADVDDQLKRIFKQLGSPSEDNWPSITQLPDYKPYPIYHPTLTWSQ 250
251 IVPNLNSRGRDLLQKLLVCNPAGRIDADAALRHAYFADTSDV         292
        10        20        30        40        50
```

Fig 17. PSPIRED V4.0 annotation grid for CDK5 C. elegans.

14 helix sections are identified(pink) and 8 strand sections are also identified with coil sections between.



pGenTHREADER Table

Modeller Licence Key:

Filter Table Rows

Min P-Value:
Max P-Value:

Show 10 entries

Search:

Conf.	Net Score	p-value	PairE	SolvE	Aln Score	Aln Length	Target Len	Query Len	Fold	SCOP	CATH	PDBSUM	Alignment	MODEL
CERT	195.674	9e-19	-340.2	-21.1	1117.5	275	282	292	6q4gA0	SEARCH	SEARCH	Open	View	Model
CERT	189.253	4e-18	-330.4	-15.4	1084.9	288	294	292	6gu2A0	SEARCH	SEARCH	Open	View	Model
CERT	184.943	1e-17	-289.0	-13.6	1069.5	285	308	292	5g6vA0	SEARCH	SEARCH	Open	View	Model
CERT	176.099	8e-17	-303.6	-14.6	1002.0	286	296	292	5efqA0	SEARCH	SEARCH	Open	View	Model
CERT	175.314	1e-16	-361.6	-15.5	982.2	277	277	292	1ob3A0	SEARCH	SEARCH	Open	View	Model
CERT	169.625	4e-16	-296.4	-15.3	957.3	287	294	292	4krdA0	SEARCH	SEARCH	Open	View	Model
CERT	164.699	1e-15	-306.4	-16.9	919.1	275	276	292	4au8A0	SEARCH	SEARCH	Open	View	Model
CERT	164.390	1e-15	-263.4	-15.4	932.7	264	284	292	5acbC0	SEARCH	SEARCH	Open	View	Model
CERT	161.440	3e-15	-313.5	-11.1	905.9	281	288	292	3rizA0	SEARCH	SEARCH	Open	View	Model
CERT	159.996	4e-15	-321.2	-18.2	881.0	271	274	292	3gbzA0	SEARCH	SEARCH	Open	View	Model

Showing 1 to 10 of 447 entries

Fig 18. PSPIRED pGenTHREADER Table of results. Sorted by Net Score

pGenTHREADER generated 447 predicted models with the best having a p-value of 9e-10. However, it is to point out that this is shorter than the target length so still not optimum and perfect result.

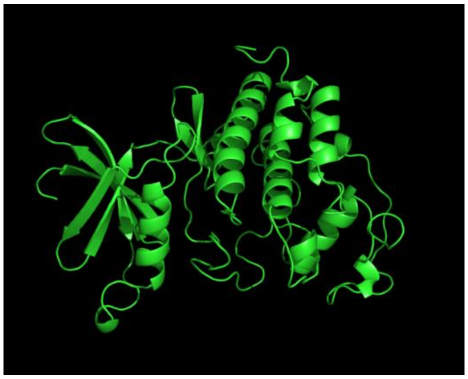
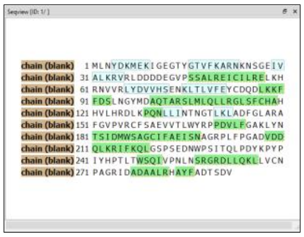


Fig 19. Best model predicted by PSIPRED pGenTHREADER



Seqview [ID: 1/]

```
chain (blank)   1 MLNYDKMEKIGEGTYGTVFKARNKNSGEIV
chain (blank)  31 ALKRVRLDDDDEGVPSSALREICILRELKH
chain (blank)  61 RNVVRLYDVVHSENKLTLVFEYCDQDLKKF
chain (blank)  91 FDSLNGYMDAQTARSLMLQLLRGLSFCHAH
chain (blank) 121 HVLHRDLKPQNLLINTNGTLKLADFGLARA
chain (blank) 151 FGVPVRCFSAEVVTLWYRPPDVLFGAKLYN
chain (blank) 181 TSIDMWSAGCIFAEISNAGRPLFPGADVDD
chain (blank) 211 QLKRIFKQLGSPSEDNWPSITQLPDYKPYP
chain (blank) 241 IYHPTLTWSQIVPNLNSRGRDLLQKLLVCN
chain (blank) 271 PAGRIDADAALRHAYFADTSDV
```

Fig 20. Best model sequence viewed with chimeraX seqview tool. Helix (green) and strands (light blue).

In this model, Fig 19, there is 11 helix sections and 7 strand sections seen in Fig. 20. This corresponds roughly to the PSPIRED results of 14 helix sections and 8 strand sections.

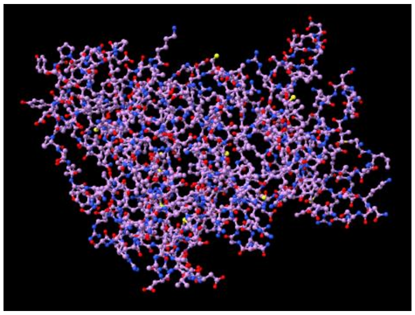


Fig 21. ChimeraX nucleotide structure prediction. Red (A), blue(T), green(G) and yellow(C).

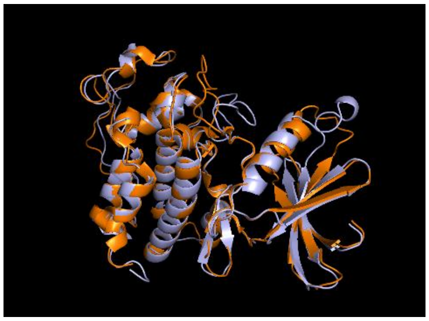
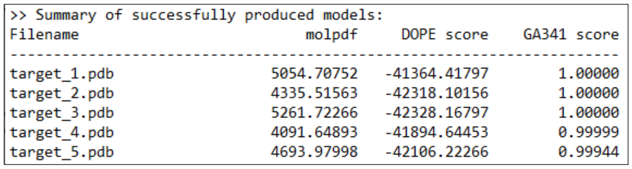


Fig 22. Best GenTHREADER model (light purple) aligned and overlayed with real model (orange).

When overlayed you can see that the start and end of the sequences are closely similar, with a lot of overlap on the helixes and strands. The differences are in the middle where no strand or helix occurs(coil).



```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
target_1.pdb                  5054.70752   -41364.41797        1.00000
target_2.pdb                  4335.51563   -42318.10156        1.00000
target_3.pdb                  5261.72266   -42328.16797        1.00000
target_4.pdb                  4091.64893   -41894.64453        0.99999
target_5.pdb                  4693.97998   -42106.22266        0.99944
```

Fig 23. Summary of Modeler results and respective DOPE scores.

Fig 23. Shows the results from the modeler log file when attempting to model the sequence from Fig. 11 – the seed shrimp. Target 1 gives lowest DOPE score of -41364.41797 and GA341 score of 1.0. The DOPE score suggests it is rather far from the real protein prediction.

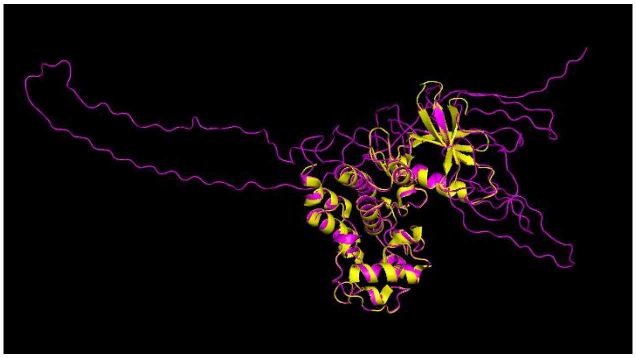


Fig 24. target_1 (pink) aligned and overlayed with real C. elegans model (yellow).

When visualized you can see the target_1 is much longer than the CDK5 protein as expected with a wild coil section to the left that is far different from the real. However, some part does align well. It is confident to say this is the protein kinase-like function's structure.

4 Discussion

The CDK5 gene in Wormbase/Ensembl has 1 transcript available with 6 exons. Out of the predicted FGENESH genes, gene two in Fig. 2 resembles the introns/extrons structure on Ensembl also matching with 6 exons. However, only exon 5 was matched identically. Interestingly, even though the exons were predicted different, it was the same protein sequence used in Swiss-Prot so FGENESH isolated the same protein, but predicted the exons different as identified by the UniProt BLAST algorithm. This 6 exon structure is further confirmed by geneid's predictions. Comparing to UniProt this is what we see in the ENSEMBL entry T27E9.3.1 confirming their results.

Homologues were identified by BLASTp with the closest relations been with other types of worms in the Caenorhabditis genus. Furthermore, a likely discovery was made that there is a gap in the Caenorhabditis bovis DNA, with a currently unnamed protein giving a match. UniProt's family domain relates heavily to Protein kinase domain. All of the 100 most similar sequences were very closely related, suggesting that this gene is homogenous to many eukaryotic organisms.

Gene function gave the same ATP binding site predicted at 33 as UniProt. Further, their exists a binding region between 10-18 on all homologues which is congruent to the UniProt annotation. There is also strong evidence that there is an ATP binding site at position 31 which is not currently annotated on UniProt. The general family the gene belongs to was identified as tyrosine and serine/threonine kinase which also corresponds to the Swiss-Prot annotation.

Protein structure was predicted with a p-value of 9e-10 which is strongly related to the existing model, although this could be improved with future work. The homologue in Darwinula stevensoni was then modeled and compared to the existing model for CDK5 in C. elegans. The kinase-like section of the protein aligns well to the CDK5 protein, suggesting that this gene function was identified and modeled correctly, within error, to that of the real structure.

Overall, this work confirmed UniProt's annotation of the 33 ATP binding site. It opened possible closely related homologues that are not annotated (Caenorhabditis bovis). It predicted a structure of Darwinula stevensoni which does not currently exist in the knowledge base, which aligned with the kinase-like region of the CDK5 in C. elegans. The work also confirmed within reason that the structure of the C. elegans protein is close to the real.

Acknowledgements

Dr. Jaume Bacardit and Dr. Simon Cockell for their teaching.

References

UniProt (2021) *Uniprot.org*.

UniProt | About (2021) *Uniprot.org*.

Flegel,S. (2021) A history of Swiss-Prot. *Sib.swiss*.

What is UniProt? | UniProt (2021) Ebi.ac.uk.

SCQ | WHAT IS BIOINFORMATICS? (2021) SCQ | The Science Creative Quarterly.

About MODELLER (2021) *Salilab.org*.

cdk-5 - Cyclin-dependent-like kinase 5 - Caenorhabditis elegans - cdk-5 gene & protein (2021) *Uniprot.org*.

UCSF ChimeraX Home Page (2021) *Rbvi.ucsf.edu*.

Gene: cdk-5 (WBGene00000407) - Summary - Caenorhabditis_elegans - Ensembl Genomes 51 (2021) *Metazoa.ensembl.org*.

cdk-5 (gene) - WormBase : Nematode Information Resource (2021) *Wormbase.org*.

BLAST: Basic Local Alignment Search Tool (2021) *Blast.ncbi.nlm.nih.gov*.

Clustal Omega < Multiple Sequence Alignment < EMBL-EBI (2021) *Ebi.ac.uk*.

Ke Chen,L. (2021) NsitePred. *Biomine.cs.vcu.edu*.

MOTIF: Searching Protein Sequence Motifs (2021) *Genome.jp*.

Bank,R. (2021) RCSB PDB - 3O0G: Crystal Structure of Cdk5:p25 in complex with an ATP analogue. *Rcsb.org*.