

Assignment 2

16006755

Word count: 2717

1 Introduction

UniProtKB (UniProt Knowledgebase) is a functional annotation database of proteins, where information is typically consistent and accurate. UniProtKB is split into two sections. "UniProtKB/Swiss-Prot" contains high quality manually-annotated records, where information is derived from literature or from peer-reviewed computational analysis. To ensure entries are consistent, the process of manually annotated notes involves six steps: sequence generation, sequence analysis, finding relevant literature, family-based curation (phylogenetics), information attribution to its original source, and quality control of completed entries. (Famiglietti ML, *et al.*, 2014)

"UniProtKB/TrEMBL" comprises of unreviewed protein sequences which are automatically computationally characterised and annotated (The UniProt Consortium). Automatic annotation involves using InterPro, a protein family classification tool, to predict presence of functional domains and significant sites. InterPro incorporates protein function prediction models from an array of databases such as PROSITE, Pfam and SMART.

TrEMBL was introduced following an increasing flow of data from experiments as Swiss-Prot's labour and time intensiveness prevented its expansion to every available protein sequence. The protein entries stored in TrEMBL are kept separate from the entries in Swiss to avoid dilution of high-quality data (The UniProt Consortium).

Over 95% of protein sequences present in UniProtKB come from the PDB database or are identified from coding sequences (CDS) translations, which are subsequently uploaded to the EMBK-Bank/GenBank/DBJ databases (INSDC). The coding sequences are produced by gene predication algorithms or have been proved experimentally. Each translated CDS is assigned a unique protein identifier (UPI) to be universally across each database to avoid redundancy between databases (The UniProt Consortium).

Neuropeptides are short amino acid sequences which modulate synaptic activity. Most neuropeptides in *C. Elegans* fall into two families: insulin-like peptides (ILPs) (Pierce *et al.*, 2001; Li *et al.*, 2003), and FMRFamide-related peptides (commonly referred as FLPs in *C. Elegans*) (Li *et al.*, 1998; Li, 2005). Most *flp* genes (*flp-1* to *flp-23* and *flp-28*) were identified from cDNA isolation and performing BLAST searches. *Flp-27* was identified by EST data mining. FLP-27, along with most FLP proteins, has been shown to be involved many processes, including egg laying, movement and fat storage (Nelson, *et al.*, 1998), from biochemical analysis involving isolation of deletion mutant FLP proteins.

The status of FMRFamide-like neuropeptides 27 (FLP-27) in UniProt is reviewed (Swiss-Prot) and the annotation score indicates there is experimental evidence at the protein level. This type of evidence supports the existence of the protein but does not shed light on the accuracy of the protein sequence displayed. Therefore, comparing attributes of the protein described in UniProt with other databases, such as WormBase would provide further insight to the veracity of the UniProt database.

2 Methods

2.1 BLAST

BLAST (basic local alignment tool) is used to search for similar sequences in its database for the provided query sequence. After saving the FLP-27 protein FASTA sequence from UniProt, protein BLAST searches were run against UniProtKB - to determine if any additional sequences are generated from the same gene, and to identify homologs. Differences between the UniProt report and the blast search are shown (Figures 5, 6). The E value corresponds to the number of expected hits of similar score that could be found by chance, thus the lower the e-value the better the match.

2.2 Ensembl

FLP-27 transcript information, including exon sequences, positions and lengths, was obtained in EnsemblMetazoa and compared against the UniProt record. Orthologues of FLP-27 were also displayed in Ensembl, which were subsequently cross-referenced with BLAST to verify them.

2.3 FGENESH

FGENESH is a gene prediction program and was used to identify exons from *flp-27* nucleotide sequence. The generated exons and sequence information was compared with the Ensembl record and discrepancies between the two sources are shown in Figure 9.

2.4 PSI-BLAST

To find functionally and evolutionary related proteins, PSI-BLAST (Position-Specific Iterated BLAST), was used. PSI-BLAST uses a sensitive profile-searching technique to

search for functional protein homologues. This method is superior than normal BLAST for detecting distantly related sequences to the query sequence as simple pair-wise sequence comparison only detects small proportions of distant evolutionary relationships. Ideally, comparing three-dimensional structures of proteins would display conserved relationships, but all proteins do not have an available 3D structure. The results from the PSI-BLAST search are noted in Figure 7.

2.5 Multiple-Sequence Alignment

Multiple-Sequence Alignment (MSA) was performed using TCOFFEE and Clustal Omega, and conserved protein motifs were displayed in JalView. MSA was used to highlight areas of similarity between FLP-27 (*C. Elegans*), CRE-FLP-27 (*C. Remanei*), FL83_04387 (*C. Latens*), B9Z55_006303 (*C. Nigoni*) and CBR-FLP-27 (*C. Briggsae*) for further phylogenetic analysis.

2.6 Gene Expression Experiments

Searching scientific literature databases such as PubMed for experimental findings about FLP-27 were performed using ArrayExpress, ExpressionAtlas and Gene Expression Omnibus (GEO). Relevant obtained literary findings were compared with the various protein analysis results. From these scientific journals, important annotation such as names of genes and proteins, protein function and sub-cellular localization were noted and cross-referenced with the UniProt database.

2.7 BUSCA

BUSCA (Bologna Unified Subcellular Component Annotator) is a protein subcellular localisation tool. This was used to verify FLP-27's active location in extracellular space. BUSCA also provides predictions of protein features, which were used to compare with the UniProt record.

2.8 3D Model

The 3D structure of FLP-27, proposed in UniProt, is generated from the Swiss-Model Repository (SMR). As the protein structure is predictive, a 3D model of FLP-27 was generated in LOMETS to be compared with the SMR model. The quality of the models produced by SMR are represented by a QMEAN (Benkert et al.). This functions as an estimator for differing geometrical properties, providing global (entire structure) and local (per residue) quality estimates for the model. The white area in each bar plot represents the protein's properties are comparable to experimentally derived proteins of similar size. If the score is positive, the model scores higher (on average) than experimental structures. If the score is negative, the model scores lower than experimental structures.

LOMETS (Local Meta-Threading Server) uses a meta-threading method to predict protein structures based on templates. The models produced by LOMETS are ranked based on Z-score, where a Z-score >1 indicates good alignment.

2.9 Protein-interaction Network

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) was used to generate a protein-interaction map centered around FLP-27. The sources and level of certainty for each interaction was noted and described in Figure 12. Identifying FLP-27 interactive partners would yield greater insight to the function of the protein, to be compared with the UniProt record. Evidence suggesting functional links between the proteins in the network are indicated by the combined score (co-expression and co-mention in PubMed) of the interactions.

2.10 Phylogenetic Analysis

In the UniProt record for FLP-27, several phylogenetic trees are linked under "Family & Domains". To verify the accuracy of these trees, comparing the listed trees with generated trees using Clustal Omega and TCOFFEE (using FASTA sequences from FLP-27 (*C. Elegans*), CRE-FLP-27 (*C. Remanei*), FL83_04387 (*C. Latens*), B9Z55_006303 (*C. Nigoni*) and CBR-FLP-27 (*C. Briggsae*).

3 Results and Discussion

3.1 Sequence accuracy

The UniProt record of FLP-27 suggests the sequence length of the protein is 89 AA (Figure 1.). This was confirmed to be correct as the amino acids were counted to be 89. Furthermore, to cross-reference this information, under "Transcript table" of FLP-27 in EnsemblMetazoa, and under "Sequences" in WormBase, the protein length is shown to be 89aa (Figure 2.).

The amino acids in FLP-27 were cross-referenced with RefSeq database, WormBase and EnsemblMetazoa and were all the same, suggesting the accuracy of the UniProt record for protein sequence is high (Figure 3.).

The nucleotide sequence of the flp-27 genomic segment was retrieved in EnsemblMetazoa, along with the upstream and downstream sequence. The sequence was uploaded to a web-based prediction program, FGENESH, to predict gene structure and to compare with the ensembl record. Interestingly, the predicted mRNA and protein in FGENESH was different to the one suggested in Ensembl, where the mRNA was shown to be 297bp and the protein to be 98 AA in FGENESH, in contrast to the 567 bp mRNA and 89 AA protein in Ensembl (Figure 9). Starting positions, end positions and lengths of the 3 exons in the protein coding transcript in Ensembl are displayed: (Figure 9 (1))

Exon 1: 5,680,980 > 5,681,107, 128
Exon 2: 5,681,851 > 5,681,954, 104
Exon 3: 5,682,333 > 5,682,667, 335

Exon positions are lengths generated from FGENESH are shown: (Figure 9 (2))

Exon 1: 1228 > 1293 (1294), 66
Exon 2: 1840 (1842) > 1973 (1974), 132
Exon 3: 2354 (2355) > 2447, 93

FGENESH correctly identified three exons. The exon lengths are different however, which could be attributed to skipping of the first Ensembl exon, only identifying the second and third Ensembl exons. To see if this is the case the DNA sequences were compared. The sequence in FGENESH starts at “ATGATTC...” which does not match the first exon sequence highlighted in blue in Ensembl (starts with “ATGTTCT...”). The second and third exons in Ensembl (starting “CCAATC...” and “ATTCCG...”) are identical to the sequence proposed by FGENESH. The causes of these discrepancies may be attributed to alternative splicing, frameshift mutations, erroneous initiation sites, natural variations and incorrect exon boundaries.

3.2 BLAST

UniProt suggests the protein is involved in a neuropeptide signaling pathway. A blastp search was used to assess this. The BLAST report suggested that the query protein is most likely a neuropeptide, a neuronal signaling protein used to directly or indirectly modulate synaptic activity (Figure 4.). As the function of FLP-27 was electronically annotated, searching for relevant transcriptomics experiments involving flp-27 would provide greater insight to the real function of the protein. The gene was searched for in ArrayExpress and ExpressionAtlas where no gene expression experiments were found. However, searching for flp-27 in GEO (Gene Expression Omnibus) provided 30 results. Since Flp-27 has been confirmed as a neuropeptide (Li et al., 1998; Li, 2005) and is therefore involved in many processes, such as locomotion, egg-laying and stress responses (Li, et al., 2014), updating the UniProt record to include how FLP-27 specifically inhibits neuronal activity (in most behaviours tested by Li) may be beneficial.

BLASTing the FLP-27 AA sequence search also showed that there are no conserved domains for this protein. To cross-reference this, NCBI conserved domain search of FLP-27 was used, where the results confirmed there were no conserved domains (Figure 5.). However, from a 2008 study by Li (Li, et al., 2008), mechanisms to process FLP precursor proteins (at the C-terminus) were similar to mammalian processes, suggesting the C-terminus may be conserved. Furthermore, a 2019 study by Gershkovich, pharmacological and function similarities have been shown between the human neuropeptide system and *C. Elegans* (Gershkovich, et al., 2019). Human neuropeptides were demonstrated to functionally compensate for *C. Elegans* neuropeptides, thereby suggesting an orthologous relationship. This is further supported by the similarity of neuropeptide requirement for receptor activation between the two species. Further analysis into the exact mechanism of action and molecular studies with neuropeptide receptors would provide greater insight into the homology of FLP neuropeptides.

The UniProt record also states that FLP-27 comes from *Caenorhabditis elegans*. This annotation also was investigated using BLASTp. The search results confirmed *C. Elegans* as the most likely organism as the BLAST hit with the lowest E.

Value and highest percentage identity came from FLP-27 [*Caenorhabditis elegans*]. The nearest hit from *C. Elegans* was from *Caenorhabditis Remanei*, where the e-value was $3e-32$. As the e-value < $e-10$ the *C. Elegans* and *C. Remanei* hits are most likely to be homologous. The top 5 hits in the BLAST search with e-values < $e-10$ are shown in Figure 6, where all are therefore suggested to be homologous.

To further assess the evolutionary relationships of FLP-27, a PSI-BLAST search of the FLP-27 sequence was used. From the first run, the e-values between FLP27_CAEL and FLP27_CAEBR are < $e-10$ suggesting an evolutionary relationship between the two proteins. Interestingly, new sequences were detected from the second round of the search, which adds the e-values from the first round to the alignment profile, suggesting the FMRFamide-like neuropeptide 27 from *Caenorhabditis Briggsae* is most evolutionary related to FLP27_CAEL.

From the BLASTp search of FLP-27, protein sequences (FASTA) from EASAFGDIIGELKGKGLGGRMRFamide [*Caenorhabditis elegans*], CRE-FLP-27 protein [*Caenorhabditis remanei*], hypothetical protein FL83_04387 partial [*Caenorhabditis Latens*], hypothetical protein B9Z55_006303 [*Caenorhabditis Nigoni*] and CBR-FLP-27 protein [*Caenorhabditis Briggsae*] were aligned using Clustal Omega and TCOFFEE to examine sequence alignment (Figure 8.). Conserved protein motifs are displayed in the consensus sequence and conservation sequence, corresponding to the (absent) functional domains identified by BLAST. The alignment between TCOFFEE and Clustal omega in JalView have generally the same shape, however the length in Clustal Omega is longer (approx. 4 amino acids). The consensus pattern is also generally the same, where slight differences are shown to occur.

FLP-27 was shown to have 4 orthologues in Ensembl (Figure 14). Cross-referencing with BLAST confirmed there are 4 orthologs (Figure 14).

3.3 Protein Localisation

The location and topology of the mature protein in the cell from UniProt is suggested to be secreted, located outside cell membranes. Using BUSCA, a protein subcellular localisation predication tool developed in 2018, the protein was suggested to be located in “Extracellular space”, which is consistent with the UniProt record. Furthermore, BUSCA suggests FLP-27 is a signal peptide, which is consistent with all databases tested.

3.4 3D Model

UniProt provides the 3D structure of FLP-27 from Swiss-Model Repository (SMR) and ModBase. The amino acid sequence of FLP-27 has been unsuccessfully processed in ModBase, but an interactive model has been produced by SMR.

The Local Quality Plot (Figure 10 (3)), from the SMR model, demonstrates the expected similarity to the native structure for every residue. Scores below 0.6 indicate the model is of low quality. The Comparison Plot (Figure 10 (4)) plots the protein length against the normalised QMEAN score, thus

comparing model quality scores to scores obtained for experimental structures of similar size.

QMEAN is shown to lie in the red region, with a score of -1.76 (Figure 10 (2).), indicating the model has below average quality. The local Quality Plot and Comparison Plot shown in Figure 10 (1, 2) also suggest the quality of the model is low.

Since the model proposed by SMR was of below average quality, LOMETS, a meta-threading method for template-based protein structure prediction, was used to generate a 3D model of FLP-27 to be compared against the SMR model. The top-ranking model (Figure 11 (1)) has Norm. Z-score of 0.55, indicating the model is of low quality.

Both SMR and LOMETS show the 3D structure of FLP-27 as an alpha helix, which is supported by the NMR analysis of *C. Elegans* neuropeptides study by (Dossey, *et al.*, 2006)

As the models from LOMETS and SMR are predictive, UniProt should consider linking the top 5 models for FLP-27 instead of just the model proposed by SMR. This would give more representative results and clearly demonstrate that the 3D structure has not been fully solved for this protein. Perhaps crystallography studies with FLP-27 would outline its protein structure more clearly, to be included in the UniProt record.

3.5 Phylogenetic Analysis

EggNOG phylogenetic tree (UniProt) listed in Figure 13 (1), along with trees produced from Clustal Omega and TCOFFEE. The trees are all shown to be the same, suggesting the accuracy for these in the UniProt record is high between each database. However, further research into phylogeny between FLP neuropeptides and human neuropeptides, as suggested by Gershkovich, may require the phylogenetic trees to be updated (Gershkovich, *et al.*, 2019).

3.6 Protein-Interaction Network

In the UniProt record for FLP-27, under “Interaction” a protein-protein interaction map for FLP-27 provided by STRING is listed (Figure 12). FLP-27 is shown to interact with other FMRFamide-like peptides (FLP-24, FLP-12, FLP-32, FLP-28 and FLP-16). Combined scores between the interacting partners are all greater than or equal to 0.633 suggesting a functional link. Annotations from STRING suggest that FLP-27 is involved in the neuropeptide signaling pathway, operating in the extracellular regions of *C. Elegans* (Figure 12), which is consistent with the UniProt record.

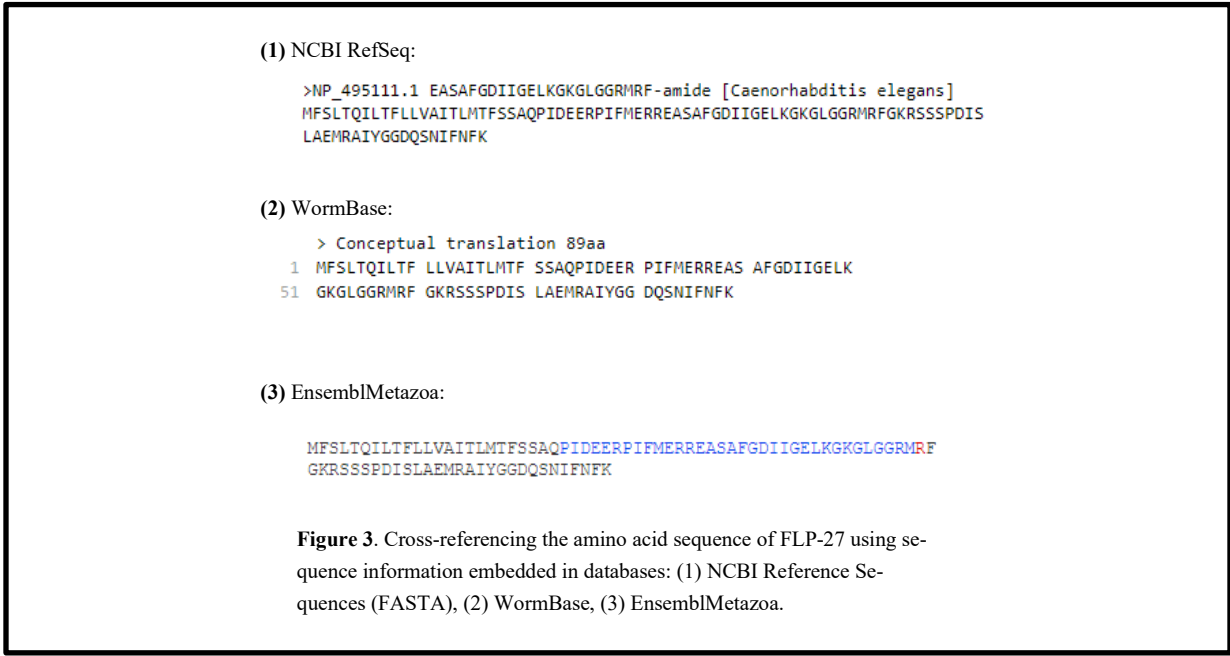
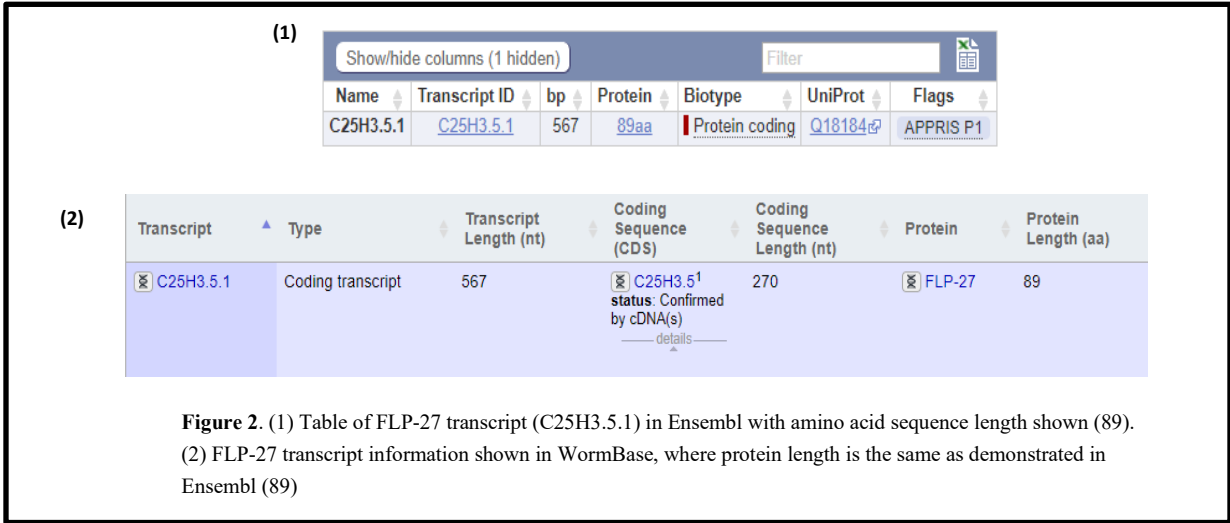
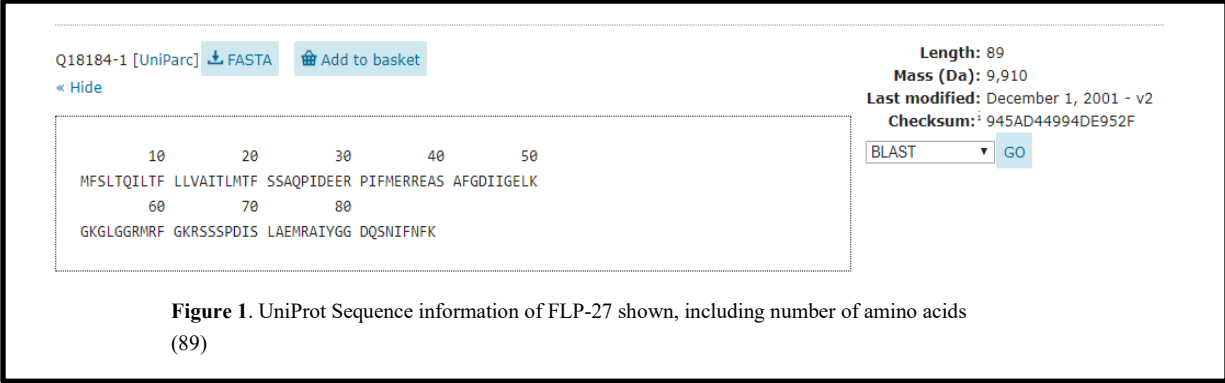
Perhaps including description about protein tertiary structure and interactions with FLP-24, 12, 16, 28 and 32 would be a useful addition to the UniProt record.

UniProt record. Updating the record with the suggestions in this paper may be beneficial.

4 Conclusion

After analysis of FLP-27 using a variety of bioinformatics tools and databases, a few discrepancies are shown in the

Figures



(1) **Functionⁱ**

FMRFamides and FMRFamide-like peptides are neuropeptides. Curated

GO - Biological processⁱ

- neuropeptide signaling pathway Source: UniProtKB-KW

[Complete GO annotation on QuickGO ...](#)

Keywordsⁱ

Molecular function	Neuropeptide
--------------------	--------------

(2) **EASAFGDIIGELKGKGLGGRMRF-amide [Caenorhabditis elegans]**

Sequence ID: [NP_495111.1](#) Length: 89 Number of Matches: 1

[See 2 more title\(s\) ▾](#)

RecName: Full=FMRFamide-like neuropeptides 27; Contains: RecName: Full=EASAFGDIIGELKGKGLGGRMRF-amide;

Sequence ID: [Q18184.2](#)

EASAFGDIIGELKGKGLGGRMRF-amide [Caenorhabditis elegans]

Sequence ID: [CCD65048.1](#)

Figure 4. (1) Protein function described in UniProt. **(2)** Results from a blastp search using FLP-27 protein sequence, with protein function highlighted under RecName.

(1) **mRNA and Protein(s)**

1. [NM_062710.6](#) → [NP_495111.1](#) EASAFGDIIGELKGKGLGGRMRF-amide [Caenorhabditis elegans]

[See identical proteins and their annotated locations for NP_495111.1](#)

Status: REVIEWED

UniProtKB/Swiss-Prot [Q18184](#)

(2) **Conserved domains on [sp|Q18184]** View [Concise Results](#) ▾

FLP27_CAEEEL FMRFamide-like neuropeptides 27 OS=Caenorhabditis elegans OX=6239 GN=flp-27 PE=1 SV=2

Graphical summary ☐ Zoom to residue level [show extra options ▸](#)


Query seq. ↑ 1 15 30 45 60 75 89
MFSLTQILTFLLVAITLMTFSSAQPIDEERPIFMERREASAFGDIIGELKGKGLGGRMRFGKRSSPDISLAEMRAIYGGDQSNIFNFK
 ... No conserved domains have been identified for this query sequence ...

[Search for similar domain architectures](#) 2 [Refine search](#) 2

List of domain hits

Name	Accession	Description	Interval	E-value

Figure 5. (1) BLASTp search of FLP-27 protein sequence, conserved domains are shown to be absent under “gene” under “Related Information” of the *C. Elegans* hit. **(2)** NCBI conserved domain search of FLP-27 protein sequence showing no conserved domains for the query sequence.



- (1) **Protein** | **FMRFamide-like neuropeptides 27**
Gene | **flp-27**
Organism | *Caenorhabditis elegans*
Status |  Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein levelⁱ

(2)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	EASAFGDIIGELKGKGLGGRMRF-amide [Caenorhabditis elegans]	178	178	100%	1e-56	100.00%	NP_495111.1
<input checked="" type="checkbox"/>	CRE-FLP-27 protein [Caenorhabditis remanei]	117	117	89%	3e-32	83.75%	XP_003108908.1
<input checked="" type="checkbox"/>	hypothetical protein FL83_04387 [Caenorhabditis latens]	116	116	89%	6e-32	82.50%	OZG21174.1
<input checked="" type="checkbox"/>	hypothetical protein CAEBREN_08869 [Caenorhabditis brenneri]	116	116	89%	6e-32	82.50%	EGT56369.1
<input checked="" type="checkbox"/>	C. briggsae CBR-FLP-27 protein [Caenorhabditis briggsae]	108	108	89%	8e-29	82.93%	XP_002630780.1

Figure 6. (1) UniProt record showing organism from which FLP-27 was derived. (2) Top 5 hits from a BLASTp search (sorted by e-value) using FLP-27 amino acid sequence.

(1)

Align. ↕	DB:ID ↕	Source ↕	Length ↕	Score (Bits) ↕	Identities % ↕	Positives % ↕	E() ↕
 1 New	SP:Q18184	FMRFamide-like neuropeptides 27 OS=Caenorhabditis elegans OX=6239 GN=flp-27 PE=1 SV=2 Cross-references and related information in: ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein sequences	89	175.0	100.0	100.0	7.0E-58
 2 New	SP:A8WU84	FMRFamide-like neuropeptides 27 OS=Caenorhabditis briggsae OX=6238 GN=flp-27 PE=3 SV=1 Cross-references and related information in: ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein expression data ► Protein sequences	90	130.0	83.0	90.0	4.0E-40

(2)



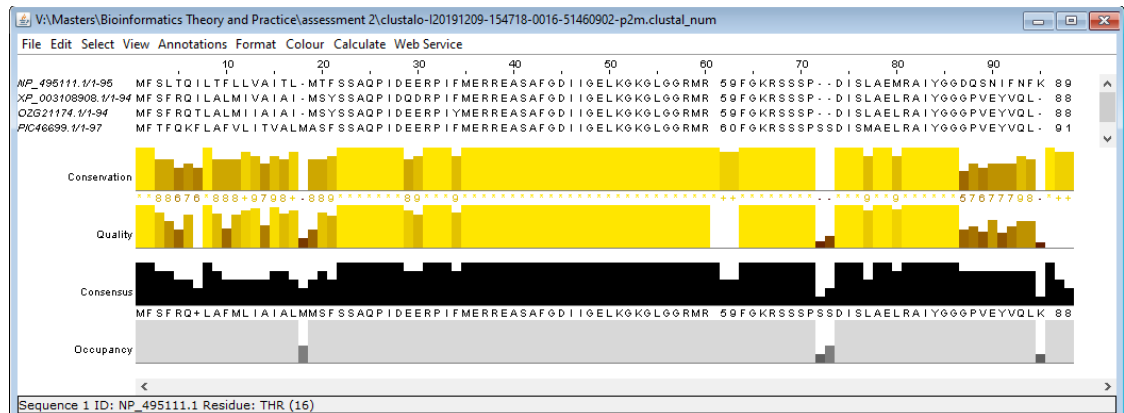
Align. ↕	DB:ID ↕	Source ↕	Length ↕	Score (Bits) ↕	Identities % ↕	Positives % ↕	E() ↕
 1 Old	SP:Q18184	FMRFamide-like neuropeptides 27 OS=Caenorhabditis elegans OX=6239 GN=flp-27 PE=1 SV=2 Cross-references and related information in: ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein sequences	89	177.0	100.0	100.0	6.0E-59
 2 Old	SP:A8WU84	FMRFamide-like neuropeptides 27 OS=Caenorhabditis briggsae OX=6238 GN=flp-27 PE=3 SV=1 Cross-references and related information in: ► Bioactive molecules ► Nucleotide sequences ► Genomes & metagenomes ► Literature ► Samples & ontologies ► Protein expression data ► Protein sequences	90	152.0	83.0	90.0	6.0E-49

Figure 7. (1) Results from the first round of PSI-BLAST, using FLP-27 amino acid sequence as the query. Two hits are shown in the first round: Alignment 1 with e-score 7.0E-58, Alignment 2 with e-score 4.0E-40. (2) Results from the second PSI-BLAST round are shown to have the same alignments as the first round with different e-scores: Alignment 1 with e-score 6.0E-59, Alignment 2 with e-score 6.0E-49.

(1)
Clustal Omega:



(2)
TCooffee:

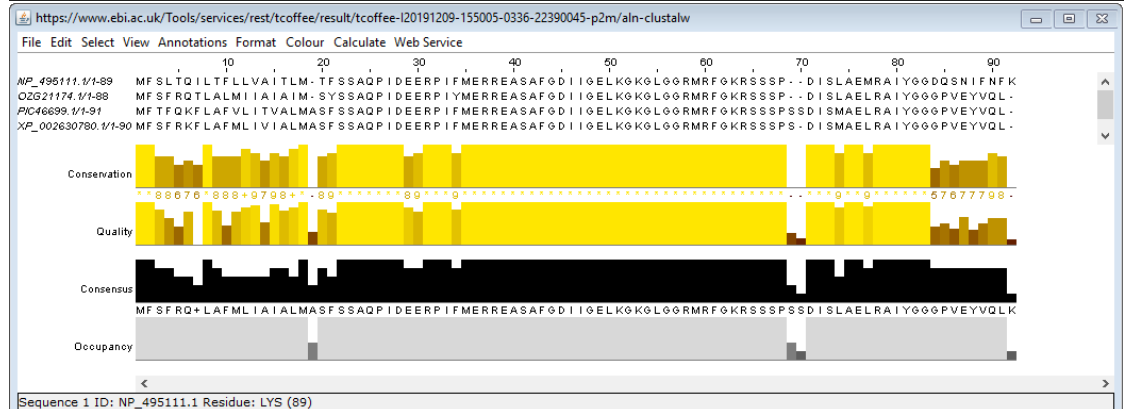


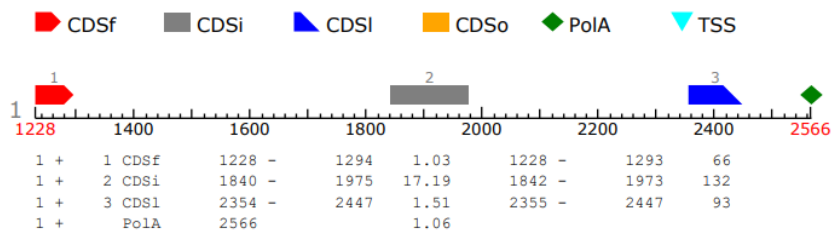
Figure 8. Multiple sequence alignment displayed in Jalview, performed using (1) Clustal Omega and (2) TCooffee. Conserved protein motifs are shown in the consensus and conservation sequence and correspond to the (absent) functional domains identified by BLAST.

(1)

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence					Cgcocctccaagtggggtgggtttcacagaggtgactgagggtttttcc
1	C25H3.5.1.e1	5 680 980	5 681 107	-	0	128	ATTGAGCTTACTTCATTTTGGTGTGCTATTCCGGTACACACAATCCTCCTCAGATGT TCTCCCTAACACAGATTCTTACTTTCTGTTGGTTCGCAATCATTGTGATGACATTCTCT CGGCTCAG
	Intron 1-2	5 681 108	5 681 850			743	gtaaaagcggatttcgggttttttagt.....tggtaatatctacatatttacag
2	C25H3.5.1.e2	5 681 851	5 681 954	0	2	104	CCAATCGACGAAGAGCGTCCGATCTTCATGGAAGTCTGTAAGCTTCAGCATTGGAGAT ATCATTGGAGAGCTTAAGSGAAAGGACTCGGCGGCGAATGAG
	Intron 2-3	5 681 955	5 682 332			378	gttgtaacctatgacaaattcatca.....ttcgaggaaaaacaaataattacag
3	C25H3.5.1.e3	5 682 333	5 682 667	2	-	335	ATTGGAAGCGATCATCTTCCCTGACATTTCAATTGGCTGAAATGCGTCAATTTATGG TGGAGACCACTCGAATACTTCACTTTAAATAATCGCAAGCTTCTGTGTTCTCCATA CCCTTCCGTTTTTCGAACATCCATCACCAGACTGATCTGTGATTGATGATCTACT TTTCATTGTTTCTCTGCGCAACCAATCAATAAAATTCAAAATTCAGAAGCGCT CCCTTTTTTCTCTTTTCATTTTGTGCTGATTTGTGTGATGATGATCATAACC GTCTTTTCTTTTTCGAAATAACAATTTTGTG
	3' downstream sequence						aagtggctcttacacatctcgaagaattgttttcaatttcattggagggga.....

(2)

FGENESH 2.6 Prediction of potential genes in *C_elegans* genomic DNA
 Seq name: II dna:chromosome
 chromosome:WBcel235:II:5679980:5683667:1
 Length of sequence: 3688
 Number of predicted genes 1: in +chain 1, in -chain 0.
 Number of predicted exons 3: in +chain 3, in -chain 0.
 Positions of predicted genes and exons: Variant 1 from 1, Score:10.210554



Predicted protein(s):

```
>FGENESH:[mRNA] 1 3 exon (s) 1228 - 2447 297 bp, chain +
ATGATTCGTAGAACTCTTTGTGCTCAGCTTACACGAAAATAAATGGAAGAGCGTAATG
TGCATTCATGATCATGGTAATATATCTACATATTTACAGCCAATCGACGAAGAGCGTCCG
ATCTTCATGGAACGTCGTGAAGCTTCAGCATTGGAGATATCATTGGAGAGCTTAAGGGA
AAGGGAATCGGCGGGCGAATGAGATTCGGAAAGCGATCATCTTCCCTGACATTTTCATTG
GCTGAAATGCGTGCAATTTATGGTGAGACCAAGTCGAATATCTCAACTTTAAATAA
>FGENESH: 1 3 exon (s) 1228 - 2447 98 aa, chain +
MIRISLCSAYTENKWSVMCIHDHGNISTYLQPIDEERPIFMERREASAFGDIIGELKG
KGLGGRMRFGKRSSSPDISLAEMRAIYGGDQSNIFNFK
```

Figure 9. (1) Exons for *flp-27* displayed in EnsemblMetazoa. Three exons are shown: Exon 1: 5,680,980 > 5,681,107 (length 128bp). Exon 2: 5,681,851 > 5,681,954 (length 104bp). Exon 3: 5,682,333 > 5,682,667 (Length 335bp)

(2) Exons for *flp-27* generated in FGENESH: Exon 1: 1228 > 1293 (1294) (Length 66bp); Exon 2: 1840 (1842) > 1973 (1974) (Length 132 bp); Exon 3: 2354 (2355) > 2447 (length 93bp)

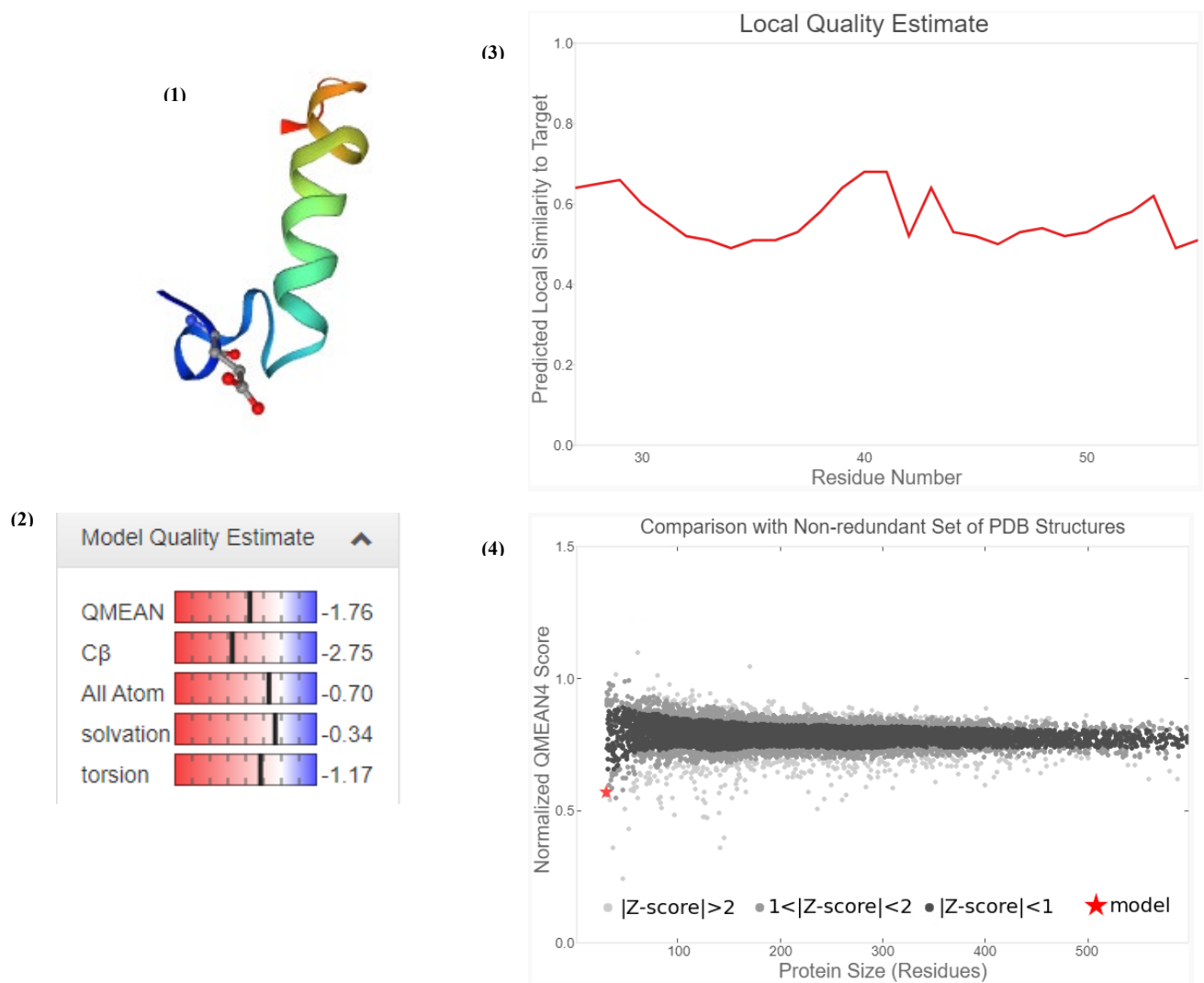
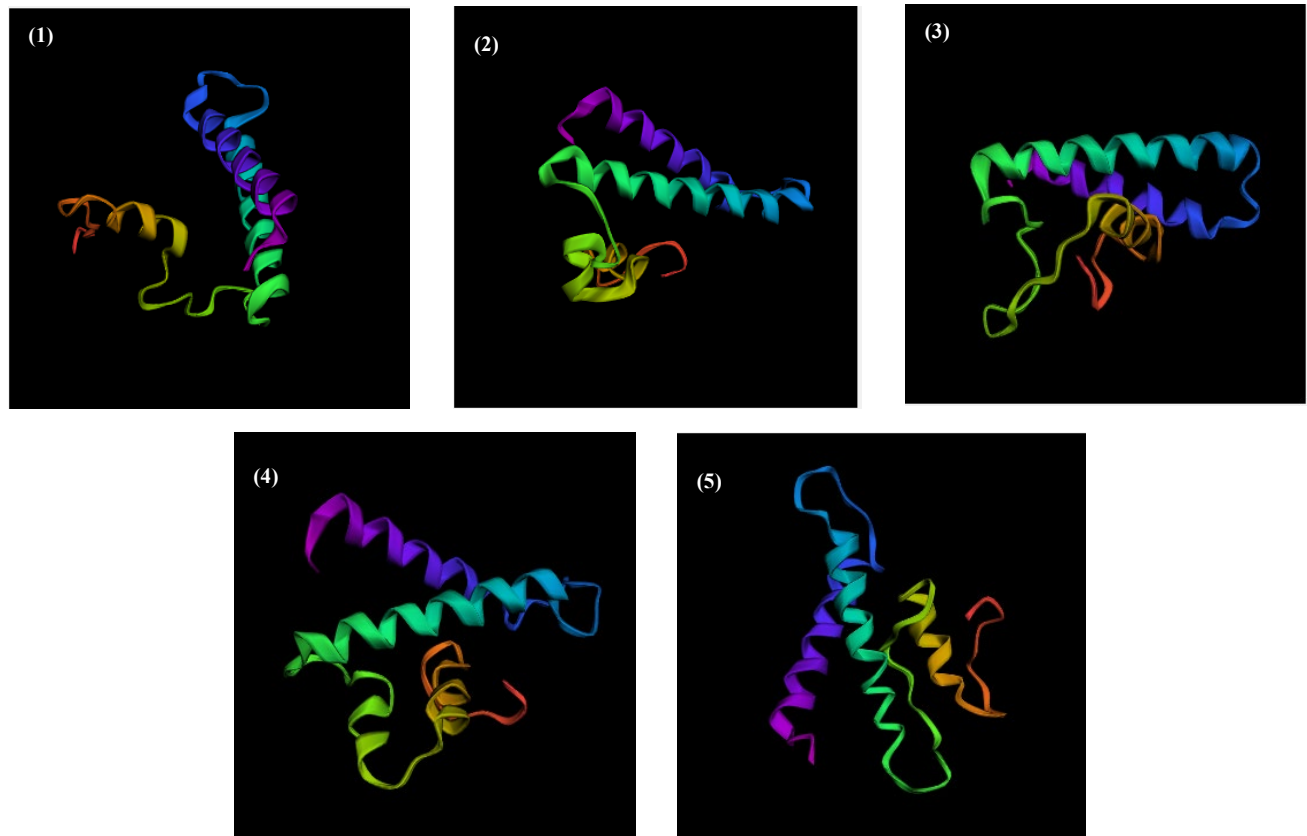


Figure 10. (1) 3D model of FLP-27 produced by Swiss-Model Repository. (2) QMEAN quality scores of the predicted protein structure. QMEAN is shown to lie in the red region, with a score of -1.76. (3) Local quality estimate for the model is shown to be ~0.6 for each residue. (4) Comparison Plot shows the model has a normalised QMEAN score of ~0.6.



(6)

Rank	PDB Hit	ID1	ID2	Cov	Norm. Zscore	Download Alignment	Gene Ontology (GO) term (Molecular Function)
1	1abrA	0.11	0.10	0.90	0.55	template_1	GO:0030598
2	2kb1A	0.07	0.07	0.92	0.54	template_2	GO:0005249
3	2zr1A	0.11	0.10	0.90	0.52	template_3	GO:0030598
4	4r8xA	0.09	0.09	1.00	0.50	template_4	
5	1orgC	0.10	0.09	0.88	0.46	template_5	GO:0005216 , GO:0005249

Figure 11. (1-5) LOMETS produced top 5 full-length models built on top templates. **(6)** Models are ranked by normalised Z-score where the top template **(1)** Top scoring model, which is shown to have Norm. Z-score of 0.55 **(6)**.

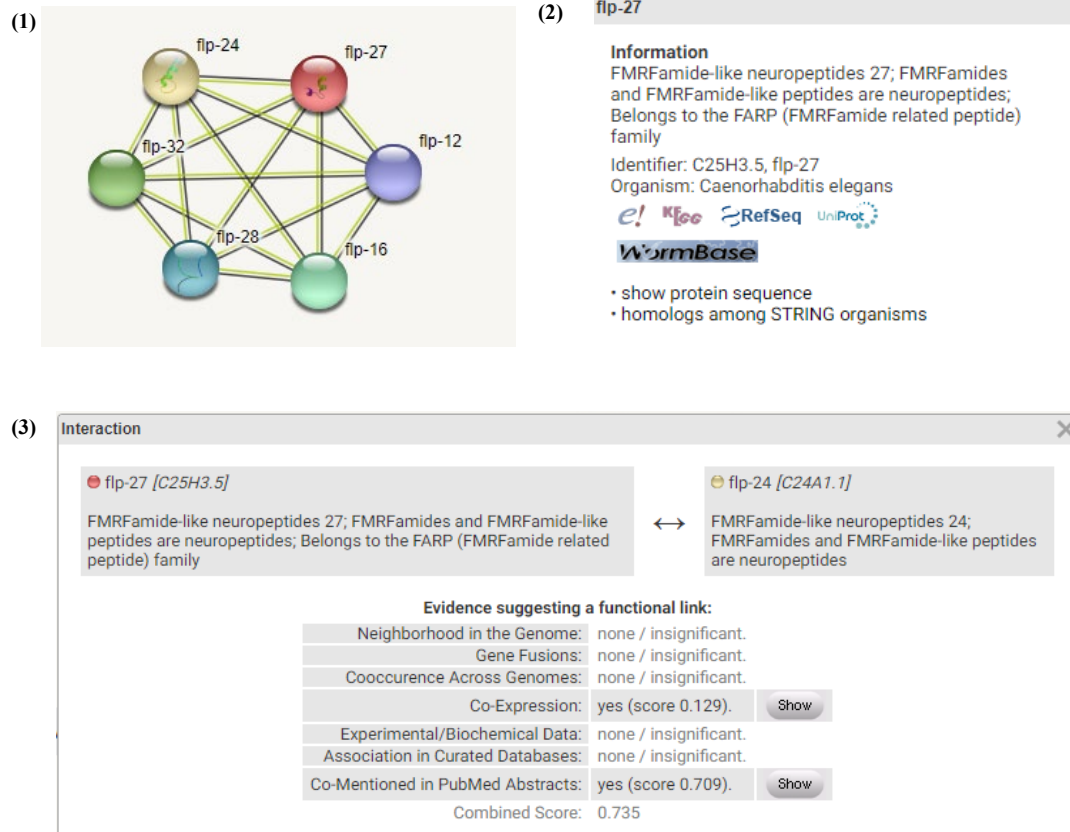
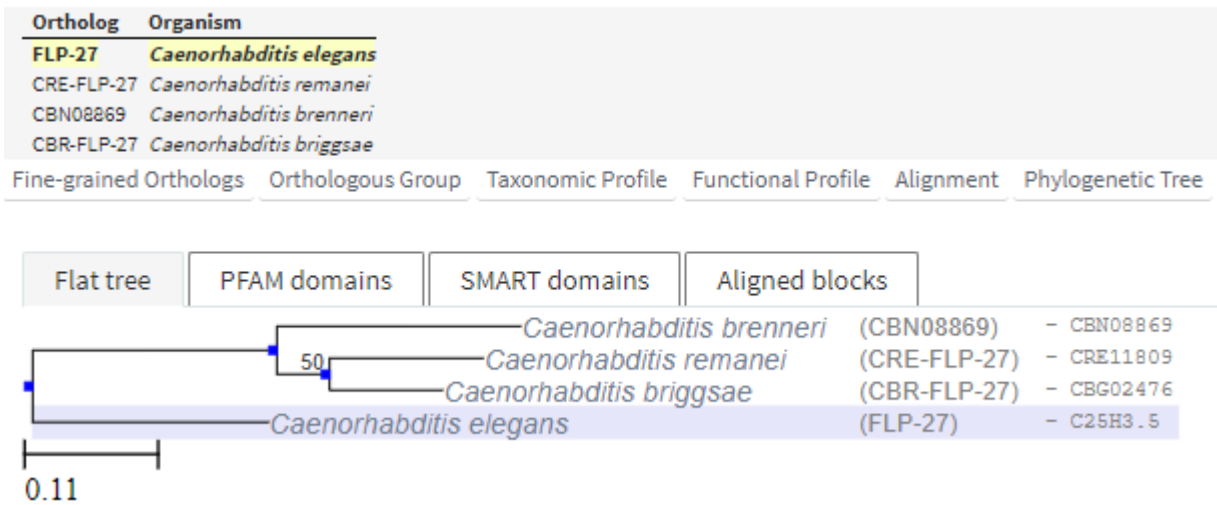


Figure 12. (1) Protein-interaction map of FLP-27 generated in STRING. (2) FLP-27 is described as an FMRFamide-like neuropeptide from *C. Elegans*. (3) Interaction between FLP-27 and FLP-24 is shown, with a combined score of 0.735 – indicating a functional link.

(1)



(2)

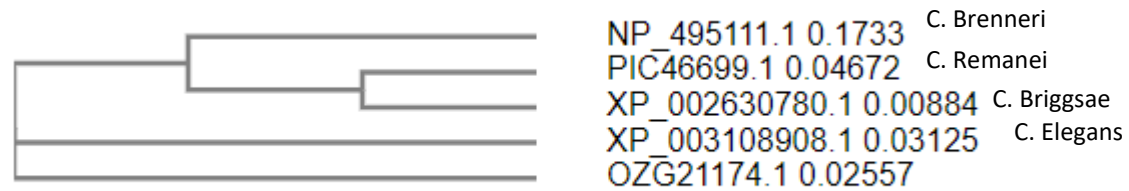


Figure 13. Phylogenetic Trees produced from (1) EggNOG, (2) Clustal Omega and Tcofee phylogenetic tree

(1)

Species	Type	Orthologue	dN/dS	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Caenorhabditis brenneri	1-to-1	CBN08869 (WBGene00147594)	0.11636	75.00 %	74.16 %	75	n/a	Yes
	View Gene Tree	Compare Regions (Cbre_Contig69:598,975-600,416-1)						
		View Sequence Alignments						
Caenorhabditis briggsae	1-to-1	Cbr-flp-27 (WBGene00025523)	0.13298	75.56 %	76.40 %	75	n/a	Yes
	View Gene Tree	Compare Regions (Il:10,033,863-10,035,426-1)						
		View Sequence Alignments						
Caenorhabditis remanei	1-to-1	Crem-flp-27 (WBGene00057307)	0.19445	76.14 %	75.28 %	75	n/a	Yes
	View Gene Tree	Compare Regions (Crem_Contig17:967,316-968,737-1)						
		View Sequence Alignments						
Schistosoma mansoni	1-to-1	Smp_070100	n/a	25.00 %	22.47 %	n/a	n/a	No
	View Gene Tree	Compare Regions (4:18,280,211-18,295,828-1)						
		View Sequence Alignments						

(2)

Group [723808at33208](#) at Metazoa level [View Fasta](#) | [View Tab Delimited](#)

C. briggsae CBR-FLP-27 protein

Functional descriptions

- GO Biological Process: 2 genes with [GO:0007218](#): neuropeptide signaling pathway
- GO Cellular Component: 2 genes with [GO:0005576](#): extracellular region

Evolutionary descriptions

- Phyletic Profile: 2 genes in 2 species (out of 450)
single copy in 2 species, multi-copy in 0 species
- Evolutionary Rate: 0.91
- Gene Architecture: Median Protein Length: 89 (std. 0.7)
Median Exon Count: 3 (std. 0)

Orthologs by organism

☐ Show all available species

Organism	Protein ID	UniProt	Description	AAs	InterPro
Caenorhabditis elegans	flp-27 (Q18184)		EASAFGDIIGELKGKGLGGRMRF-amide >>	89	

Group [1855737at2759](#) at Eukaryota level 2 genes in 2 species >>

C. briggsae CBR-FLP-27 protein

Group [17676at6231](#) at Nematoda level 2 genes in 2 species >>

C. briggsae CBR-FLP-27 protein

Group [17615at119089](#) at Chromadorea level 2 genes in 2 species >>

C. briggsae CBR-FLP-27 protein

Figure 14. (1) 4 orthologs of FLP-27 are shown in EnsemblMetazoa
(2) Cross-referencing orthologs in BLAST showed 4 orthologs of FLP-27

References

- Famiglietti ML, Estreicher A, Gos A, Bolleman J, Gehant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L, Xenarios I. Genetic Variations and Diseases in UniProtKB/Swiss-Prot: The Ins and Outs of Expert Manual Curation. *Hum. Mutat.* 35:927-935 (2014)
- The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47: D506-515 (2019)
- Pierce SB, Costa M, Wisotzkey R, Devadhar S, Homburger SA, Buchman AR, Ferguson KC, Heller J, Platt DM, Pasquinelli AA. Regulation of DAF-2 receptor signaling by human insulin and ins-1, a member of the unusually large and diverse *C. elegans* insulin gene family. *Genes Dev.* 2001;15:672-678. PMID:11274053.
- Li W, Kennedy SG, Ruvkun G. daf-28 encodes a *C. elegans* insulin superfamily member that is regulated by environmental cues and acts in the DAF-2 signaling pathway. *Genes Dev.* 2003;17:844-858. PMID: 12654727.
- Li C, Kim K, Nelson LS. FMRamide-related neuropeptide gene family in *Caenorhabditis elegans*. *Brain Res.* 1998;848:26-34. PMID: 10612695.
- Li C. The ever-expanding neuropeptide gene families in the nematode *Caenorhabditis elegans*. *Parasitology.* 2005;131(Suppl):S109-127. PMID: 16569285.
- Nelson LS, Rosoff ML, Li C. Disruption of a neuropeptide gene, *flp-1*, causes multiple behavioral defects in *Caenorhabditis elegans*. *Science.* 1998b;281:1686-90. PMID: 9733518.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.
- Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Giron, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, Paul Flicek. Ensembl 2018. PubMed PMID: 29155950
- Salamov, Asaf A., and Victor V. Solovyev. "Ab initio gene finding in *Drosophila* genomic DNA." *Genome research* 10.4 (2000): 516-522.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 25:3389-3402.
- Notredame C., Higgins D.G. and Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302(1):205-17. PubMed: 10964570
- Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. and Higgins D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. PMID: 21988835
- Athar A. et al., 2019. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* doi: 10.1093/nar/gky964, Pubmed ID 30357387.
- Petryszak R., Keays M., Tang Y.A., Fonseca N.A., Barrera E., Burdett T., Füllgrabe A., Fuentes A.M., Jupp S., Koskinen S. et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 2016; 44:D746-D752.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D991-5.
- Savojardo, Castrense et al. "BUSCA: an integrative web server to predict subcellular localization of proteins." *Nucleic acids research* vol. 46,W1 (2018): W459-W466. doi:10.1093/nar/gky320
- Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. *Nucleic Acids Res.* 45, D313-D319 (2017)
- Wu, Sitao, and Yang Zhang. "LOMETS: a local meta-threading-server for protein structure prediction." *Nucleic acids research* vol. 35,10 (2007): 3375-82. doi:10.1093/nar/gkm251
- Szklarczyk et al. *Nucleic acids research* 47.D1 (2018): D607-D613.2
- eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. *Nucl. Acids Res.* (04 January 2016) 44 (D1): D286-D293. doi: 10.1093/nar/gkv1248
- Li, Chris, and Kyuhyung Kim. "Family of FLP Peptides in *Cae-norhabditis elegans* and Related Nematodes." *Frontiers in endo-crinology* vol. 5 150. 14 Oct. 2014. doi:10.3389/fendo.2014.00150), updating the UniProt record to include how FLP-27 specifically inhibits neuronal activity (in most behaviours tested by Li
- Li, Chris, and Kyuhyung Kim. "Neuropeptides." *WormBook : the online review of C. elegans biology* 1-36. 25 Sep. 2008. doi:10.1895/wormbook.1.142.1
- Gershkovich, M.M., Groß, V.E., Kaiser, A. et al. Pharmacological and functional similarities of the human neuropeptide Y system in *C. elegans* challenges phylogenetic views on the FLP/NPR system. *Cell Commun Signal* 17, 123 (2019) doi:10.1186/s12964-019-0436-1
- Dossey, Aaron T et al. "NMR analysis of *Caenorhabditis elegans* FLP-18 neuropeptides: implications for NPR-1 activation." *Bio-chemistry* vol. 45,24 (2006): 7586-97. doi:10.1021/bi0603928