Choosing a sequence

After being assigned the 5 different protein sequences, I was unsure as to which one would be the most interesting to write about. I decided to take some time to get to know each protein before proceeding. I wanted to make sure that the protein that I chose was well annotated at the very least, as this would allow me to delve further into the analysis of the protein and its function. This led to me choosing the sequence Q0KI58 DROME (Serine/threonine protein phosphatase 2A regulatory subunit), not only was this sequence well annotated, but it also seemed very interesting. The protein sequence consists of 670 amino acids and the protein plays a role in the regulation of cell proliferation, programmed cell death, cell differentiation and embryonic development. This protein also exists in the species Drosophila melanogaster which is a common fruit fly.

Delving into the analysis of Q0KI58 DROME and its surrounding homology

To get started with, I decided that the best course of action would be verifying the integrity of the protein sequence by running the sequence through protein BLAST (nih.gov) (BLAST, n.d.). This would allow me to ensure that the sequence was in fact correct and there were no errors. The reason this is necessary is because UniProt has over 60 million sequences, it is common for errors to slip through especially if the protein isn't annotated very well. I first retrieved the protein sequence for Q0KI58 DROME from UniProt, to get this I clicked on the BLAST button on the protein sequence page which took me to a new page with the full sequence (UniProt, n.d.)

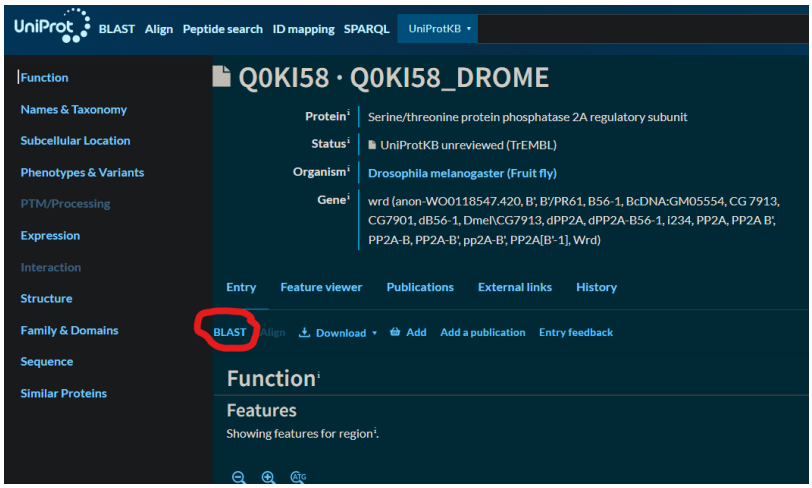

Figure 1 - Showing where the sequence is located on UniProt. (UniProt, n.d.)

I then went over to Protein BLAST (BLAST, n.d.) and pasted in the sequence under where it says, 'Enter Query Sequence', the algorithm I used was blastp. (BLAST, n.d.)

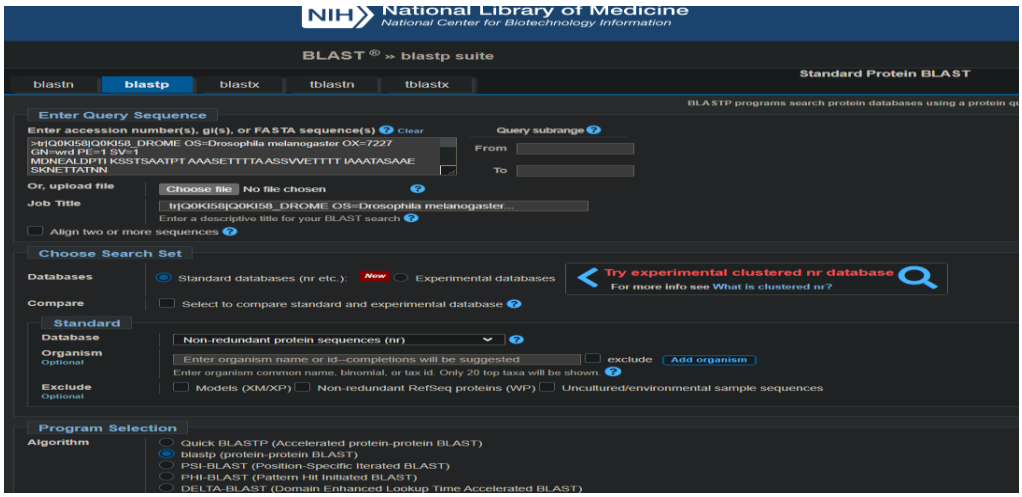

Figure 2 - Showing where to enter the sequence into on the National Library of Medicine's Protein BLAST. (BLAST, n.d.)

After querying the database using the blastp reference, there was plenty of results with over 85% similarity, and with 0.0 as their E-value. I managed to find out that Q0KI58 DROME shared a 100% similarity rate with well-rounded, isoform L [Drosophila melanogaster] (*will be referenced later*). I decided to look at the sequence entitled well-rounded, isoform P [Drosophila melanogaster]. It shared 97.84% similarity with Q0KI58 DROME. There were many similarities that suggested that the two sequences where homologous, the first being that the name well-rounded, isoform P is very similar to well-rounded, isoform L, which as previously stated returned with a 100% similarity match to Q0KI58 DROME, indicating that they all share a common ancestor. Another similarity that I picked up on upon further analysation is that both sequences shared almost identical strings of amino acids for the most part, explaining their high similarity score, however they did differentiate slightly. There was one area that showed where the query sequence (Q0KI58 DROME) differentiated from well-rounded, isoform P *(see figure 3).*

```
Query  83   TNGIKESNSNLStttaaavaaattvEGVAPAITSTIVVTGGTPPLSSLA----------  132
            TNGIKESNSNLSTTTTAAAVAAATTVEGVAPAITSTIVVTGGTPPLSSLA
Sbjct  83   TNGIKESNSNLSTTTTAAAVAAATTVEGVAPAITSTIVVTGGTPPLSSLAAQLRRQGFAQ  142

Query  133  ---NKLKDNTppydappptpISKVLNITGTPIVRKEKRQTSARYNASKNCELTALIPLNE  189
               NKLKDNTPPYDAPPPTPISKVLNITGTPIVRKEKRQTSARYNASKNCELTALIPLNE
Sbjct  143  SKANKLKDNTPPYDAPPPTPISKVLNITGTPIVRKEKRQTSARYNASKNCELTALIPLNE  202
```

Figure 3 - Taken from the 'query' (Q0KI58 DROME) protein sequence and the 'subject' (well-rounded, isoform P [Drosophila melanogaster]) protein sequence, the dashes indicate no amino acid presence. (This sequence comparison was collected by clicking on well-rounded, isoform P [Drosophila melanogaster] after receiving the results of the database from the query sequence) (BLAST, n.d.).

Another differentiation that we can infer from figure 3 regarding the two sequences is that Q0KI58 DROME consists of only 670 amino acids in total, whereas well-rounded, isoform P is made up of 683 amino acids. We can see this as if we count the number of dashes in the query sequence it adds up to 13, which shows that the subject sequence is comprised with 13 more amino acids in comparison to the query sequence.

I was interested to see how many orthologues Q0KI58 DROME had; this is essentially genes in different species that have evolved through speciation only. To do this I used EMBL-EBI's database (EMBL-EBI, n.d.) and searched 'Q0KI58'. After doing so I was able to see that the protein had 14 sequences and ontologies and 9 different protein families. This shows that Q0KI58 DROME shares a common evolutionary origin within 9 separate protein families, further validating the protein sequences' legitimacy.

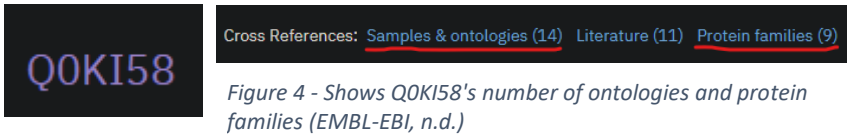

Figure 4 - Shows Q0KI58's number of ontologies and protein families (EMBL-EBI, n.d.)

Overall judging from the long list of 100 protein sequences returned when I queried Q0KI58 DROME on protein BLAST (BLAST, n.d.), and all of the sequences at least sharing a minimum of 81.45% similarity with Q0KI58 DROME's protein sequence, I can infer that most, if not all of the information surrounding Q0KI58 DROME on UniProt (UniProt, n.d.) is correct.

Q0KI58 DROME's phylogenetic tree & multiple sequence alignments

I decided to look at the assigned protein's phylogenetic tree, reason being because phylogenetic trees are an area of bioinformatics that interest me. It allows us to visualize the relationships of a certain protein sequence with other homologous sequences. This can help us create links between said protein and others, allowing us to visualize whereabouts the proteins began to differentiate from each other and split off into their own branches.

Figure 5 - Phylogenetic tree generated for Q0KI58 DROME (BLAST, n.d.)

Upon further analysis of the phylogenetic tree we can see just how closely related Q0KI58 DROME is with well-rounded, isoform P, with them being close to each other on the tree. Q0KI58 DROME is also right next to well-rounded, isoform L which as we previously found out, came back with a 100% similarity rate (explaining why they are right next to each other). Well-rounded, isoform J is also very similar to these two sequences and shares a similarity rate of 99.83% with the Q0KI58 DROME!

In order to retrieve Q0KI58 DROME's phylogenetic tree, I first pasted the protein sequence into protein BLAST (BLAST, n.d.) after retrieving it from UniProt (UniProt, n.d.). I then clicked on Multiple alignment on the other reports section of the query return page, after this I selected 5 other protein sequences, hit re-align, then clicked phylogenetic tree which is at the top right of the page. Doing the mentioned steps led to the diagram presented in figure 4 loading up.

After running the proteins listed above into EMBL-EBI's Clustal Omega multiple sequence alignment database (EMBL-EBI, n.d.), I was able to build the sequence alignment between the proteins (*see figure 6*).

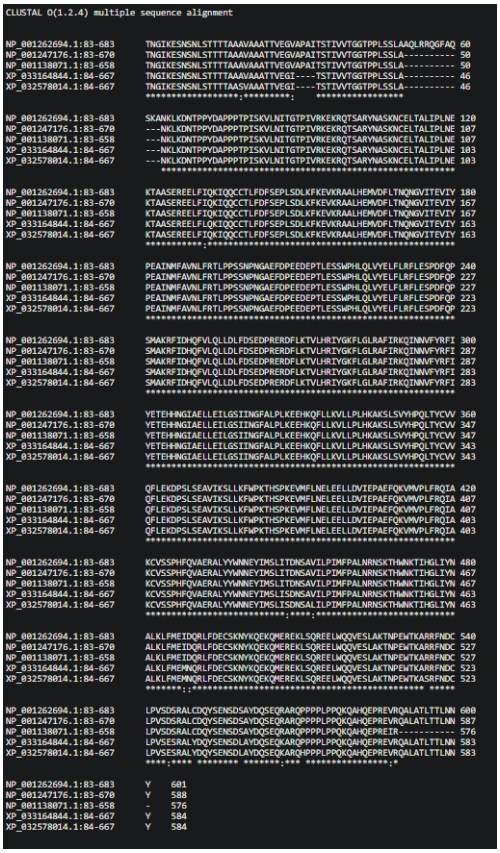

Figure 6 - The return from the multiple sequence alignment on EMBL-EBI's Clustal Omega (EMBL-EBI, n.d.)

Through figure 6 we can see clearly just how similar all the sequences truly are, there is very little differentiation overall and they seem to all consist of roughly the same number of amino acids. We can see the various evolutionary patterns and relationships emerge from this sequence alignment, it clearly demonstrates conserved regions of protein alignment across the proteins and shows once again that they share a common ancestor.

I retrieved the sequence alignment by first heading over to protein BLAST (BLAST, n.d.) and selecting all the proteins that I previously had selected for the phylogenetic tree, I selected them by ticking them. After this I downloaded them in an aligned sequence, I done this by clicking download then clicking 'FASTA (aligned sequence)'.

Figure 7 - Highlights how to download the selected sequences (BLAST, n.d.)

After this I copied the text that was within the downloaded FASTA aligned sequences file and pasted it into EMBL-EBI's Clustal Omega which can be found by just searching 'Clustal Omega' on google. I then hit submit and a few seconds later the multiple sequence alignment was ready for me to download.

Having a look at Q0KI58 DROME's secondary structure

To complete this task, I had to familiarise myself with PSPIRED (PSIPRED, n.d.), which is a method/database that is used to investigate a protein structure, it uses artificial neural network machine learning methods.

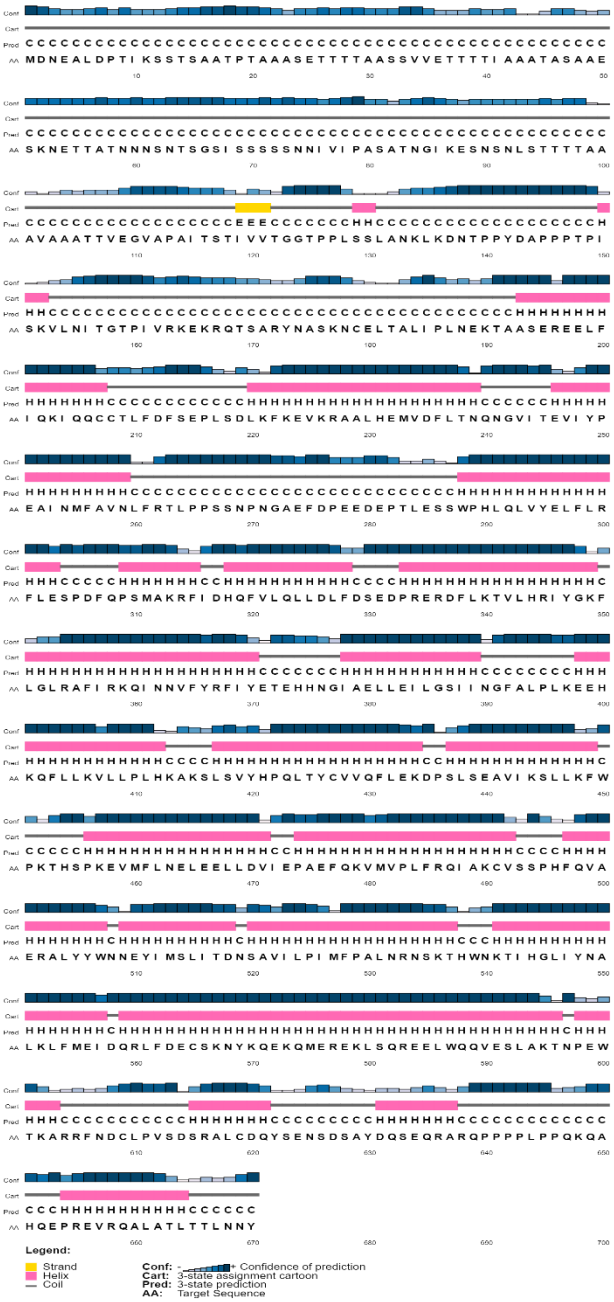

Figure 8 - PSIPRED cartoon, PSI-blast based secondary structure prediction (PSIPRED, n.d.)

This PSIPRED cartoon helps predict the secondary structure of Q0KI58 DROME, I was able to generate this cartoon and download it in PNG format from the PSIPRED website. I first obtained the BLAST sequence of Q0KI58 DROME from UniProt, I then entered it into PSIPRED's query creation sequence, after waiting 10-20 minutes the secondary structure prediction was generated. I then scrolled down and expanded the PSIPRED cartoon section, leading to the downloading of figure 8.

Gaining insight into Q0KI58 DROME's 3D model through Alphafold and SWISS-MODEL Interactive

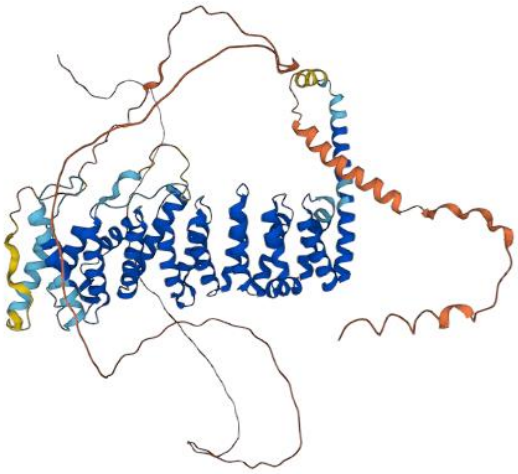

Figure 10 - 3D model of Q0KI58 DROME taken from AlphaFold (AlphaFold, n.d.)

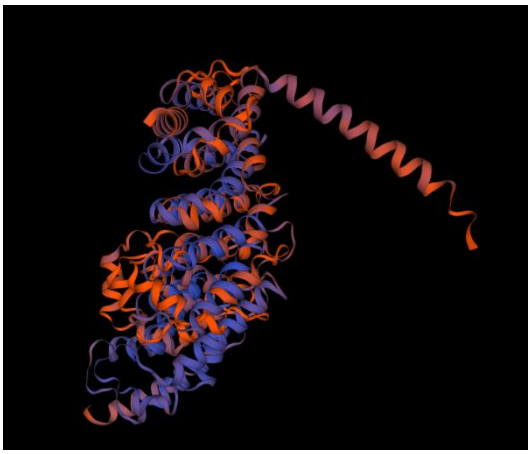

Figure 9 - 3D model of a homologue of Q0KI58 DROME taken from SWISS-MODEL Interactive (Interactive, n.d.)

Both models contain the same long string of DNA that branches off from the rest of the sequence, overall it could be said that both models look very similar, but as we can see they aren't identical. This is because the model taken from SWISS-MODEL Interactive (Interactive, n.d.) returned with a 79.85% sequence identity match when I pasted in Q0KI58 DROME's protein sequence on their website. This means that the two proteins differentiate from one another, but likely share a common ancestor and are homologous.

To retrieve the 3D model from AlphaFold (AlphaFold, n.d.), I looked at Q0KI58 DROME's UniProt page and scrolled down to the structure of the protein, I then clicked on the AlphaFold link where I was taken to AlphaFold's page designated to Q0KI58 DROME's 3D model.

To access the 3D model from SWISS-MODEL Interactive (Interactive, n.d.), I retrieved the protein sequence from UniProt then entered it into SWISS-MODEL's target sequence box. After hitting submit I had to wait 10-20 minutes for the results.

Having a look at the close network interactions of Q0KI58 DROME

BIOGRID (BIOGRID, n.d.) is an online interaction repository with data compiled through comprehensive curation efforts, essentially the website helps us to view proteins interactions with other proteins.

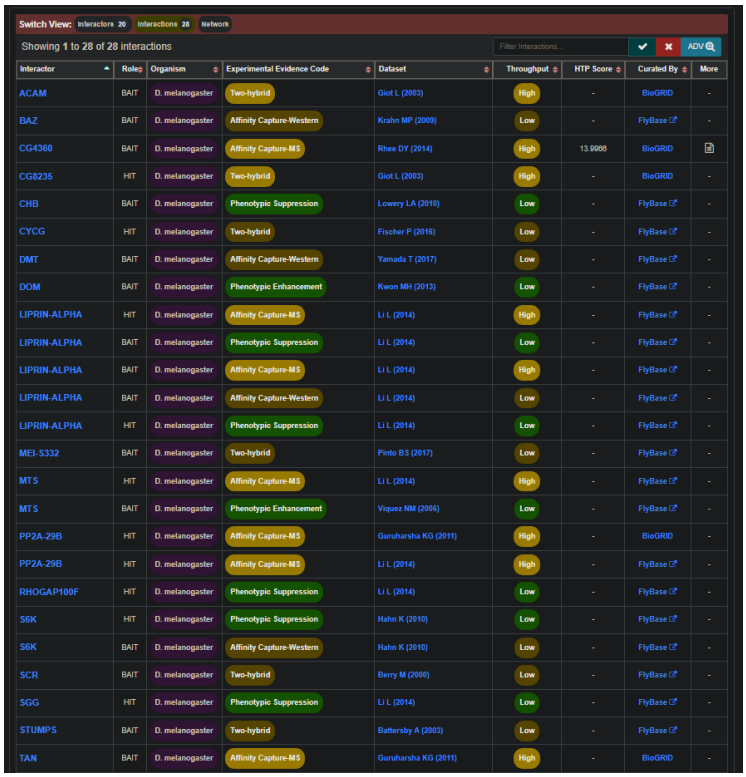

Figure 11 - Shows Q0KI58 DROME's interaction with other proteins on BIOGRID (BIOGRID, n.d.)

Here we can see various proteins that are related to Q0KI58 DROME, through the throughput, we can see how common the interaction is between Q0KI58 DROME and these numerous proteins, varying from low and high with different shades of colour.

I was able to acquire figure 11 by navigating over to BIOGRID (BIOGRID, n.d.) and searching 'Q0KI58' into their database. I then clicked on the interactions tab on the switch view taskbar to pull up the table.

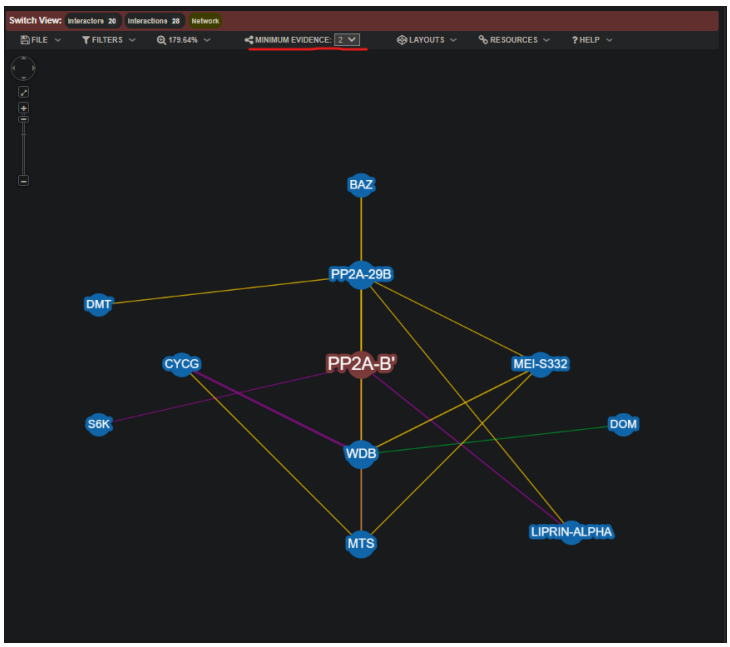

Figure 12 shows Q0KI58 DROME's network connection to other proteins in diagram form, I decided to demand more evidence by requiring 2 rather than 1, as highlighted in the screenshot. I was able to acquire figure 12 by navigating over to BIOGRID and searching 'Q0KI58' into their database. I then clicked on the network tab and changed the minimum evidence to 2.

Figure 12 - Screenshot taken from BIOGRID showing Q0KI58 DROME's network connections to other proteins (BIOGRID, n.d.)

Final thoughts on the bioinformatics of a protein

Overall I have found this module to be one of the more interesting ones, I feel as though it was very research heavy, navigating various websites and scouring databases to delve deeper into the analysis of the subjected protein. This was fun to me and a very unique experience, biology and computer science never crossed my mind as being interlinked so deeply. This section of computer science just

goes to show how broad the field of computer science truly is, it is an integral part of society that is continuing to develop and broaden as time progresses. The study of proteins in relation to bioinformatics is still a relatively new field, having only been founded in the early 1960s, yet the amount of data that is available is astronomical. It will be fascinating to see how this area of study progresses in correlation with time and how far bioengineers will progress with the study of the human genome. The program/database that I enjoyed using the most throughout this report was protein BLAST (BLAST, n.d.). The reason for this was because I found it super simple to navigate and engrossing being able to see all the protein sequences that matched the query sequence in different ways. The program that I found to be the most irritating was probably PSIPRED (PSIPRED, n.d.), this was not due to the nature of the website, more the content. I found it difficult to understand and comprehend the data that it presented.

I feel as though the information surrounding Q0KI58 DROME (Serine/threonine protein phosphatase 2A regulatory subunit) was for the most part, correct and very well annotated. It had plenty of homologous routes with other protein sequences which checked out to be correct. The UniProt (UniProt, n.d.) page was well formatted and had external links to AlphaFold (AlphaFold, n.d.) which showed the 3D structure of the protein. However, one noticeable flaw that I found within the protein page on UniProt was that the function section was empty, giving no information on the function of the protein or sequence whatsoever. I feel as though this information would've helped me to understand the protein a little bit better when getting started.

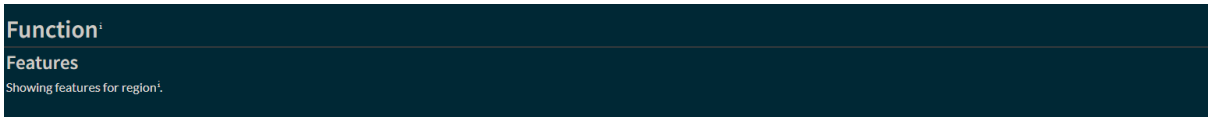

Figure 13 - Details Q0KI58 DROME's function section, highlighting its empty nature (UniProt, n.d.)