CSB352 Lab Report 1, Brandon Kim

Introduction:

In this lab report, *Arabidopsis thaliana*; a eudicot species deriving from the Brassicaceae family that is notoriously used as a model organism because of its small, completely sequenced genome, highly efficient transformation, among other things to study various mechanisms associated[2]. A protein of unknown function, with AGI ID of At5g66930, sequenced using BLASTp[3] because there is a protein query sequenced against the non-redudant protein database to output a protein alignment. The protein is 251 amino acids in length, has a molecular weight of 28297.5 g/mol, 6.95 isoelectric point, and is located on chromosome 5 between 26725839 and 26727752 base pairs with a forward orientation. A global multiple sequence alignment will be used to identify the regions of high similarity as well as observe gaps and variations in the sequences. A phylogenetic tree will also be generated to determine the evolutionary relationship clearly; to see the big picture. In this report, the function of this gene by using various methods such as the two mentioned above and using the discovered orthologs to connect a link.

Methods:

The BLASTp program was used to search for protein alignments with a protein query and a protein database. BLASTp was chosen over other algorithms because In this report, firstly, the provided gene was imputed in the TAIR[5] resource; retrieving the gene's protein sequence by inputting the accession number: At5g66930. The dataset in this website used is Araport11 protein sequences, and was searched against one sequence per locus. After retrieving this representative gene model in FASTA format, it was inputted into BLASTp with the database *non-redundant protein sequence*, with the expect threshold algorithm parameter altered to 1e-20.

The threshold algorithm was changed from 0.05 default to 1e-20 because this assures that the results are high quality, and more related to the search. In addition, in the scoring parameters, the scoring matrix was altered from the default BLOSUM62 to BLOSUM80. This is because the sequences generated will result in a higher level of sequence similarity. The gap costs was switched from the default existence 11 and extension 1 to existence 10 and existence 1 to promote a higher gap penalty, resulting in a reduce in gaps within the alignment. This will allow for some gaps, but not an abundant amount to promote better alignment. Navigating to the alignments tab will allow us to see aligned sequences as well as the comparison between the query and subject in a comprehensive way by changing the alignment view to pairwise with dots for identities. After examining this alignment view, the MSA viewer was then navigated to. With the NCBI Blast MSA import, it automatically anchors the query sequence and shows every result under it. Navigated to tools, and unchecked “show identical residues as dots”. Now, zooming in on a section will reveal all residues, by the global CLUSTAL alignment algorithm. A global algorithm was favoured over a local algorithm because the sequence is very uniformly similar across.

After alignment, we can work to build the phylogenetic tree in MEGA[4]. The aligned FASTA file was downloaded from MSA alignment on NCBI. The FASTA file was converted to a MEGA file via the conversion feature on MEGA, and the phylogenetic tree was created using this aligned sequence file. The maximum likelihood tree was created because it considers every location, with different models; making more trees. It is superior over Neighbour-joining tree in this case because we prioritized quality of tree at the cost of more time required to generate. The settings inputted are all default except for the Test of Phylogeny is set to bootstrap with 500 replications. After generating and observing the maximum likelihood tree, a neighbour-joining

tree algorithm was also created to compare the two. The only setting changed on the neighbour-joining tree algorithm is the test of phylogeny altered to Bootstrap method. Since these two phylogenetic tree methods are independent, investigating both and comparing them allows for a high level of confidence on the analysis being correct. After observing the phylogenetic tree in MEGA, the tree was also observed in Treeview[1], an alternate tree-building resource that gives a more comprehensive tree output as well as speed for generation.

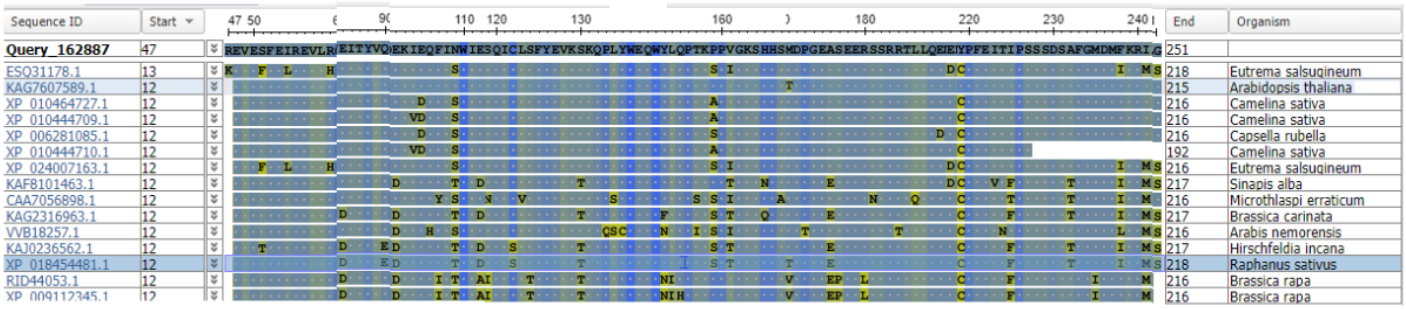
Results:

After concluding with the generation of the MSA as well as the phylogenetic trees via three different methods each, I found out several things. Firstly, the *Arabidopsis thaliana* species has a function that is shared amongst many of the closely related taxa, as seen by the percent identity and low e-values; indicating a steady, or at the least, non-negative synteny. This means that the gene is highly conserved between taxa. After generating the MSA and the phylogenetic tree, comparing the closely related species with each other and contrasting them, it can be seen that the function of the gene could possibly be protein binding; generally.

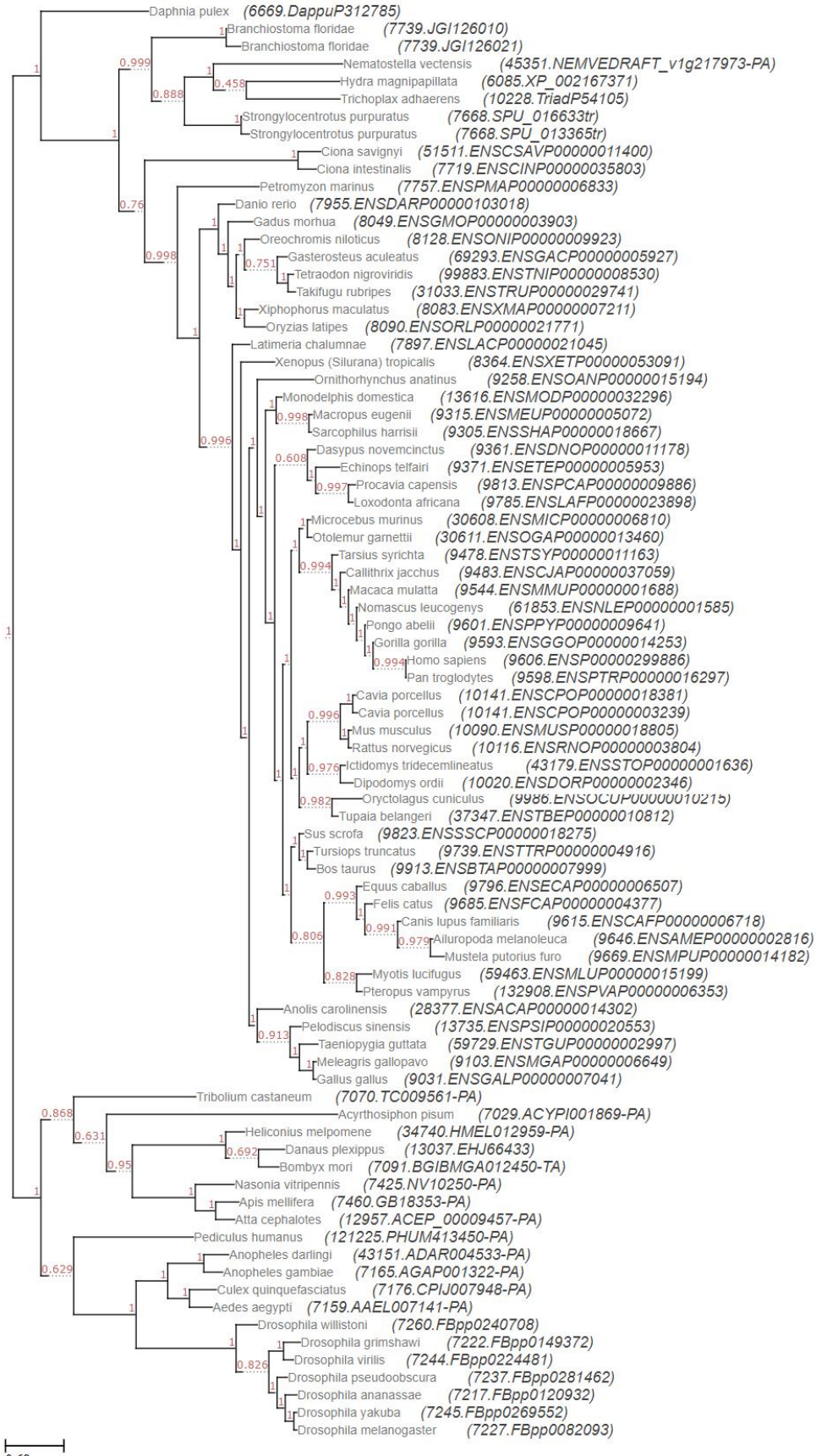
There were a few things that were observed to be out of the ordinary". When creating the Maximum Likelihood phylogenetic tree, the generation time was exceptionally long, considering the length of the FASTA file. In addition, when researching the functions of homologs and orthologs, it came to the attention that there were very similar species when the query was blasted. It was found out that a setting got changed to a value it was not supposed to, and this is why it took a long time. The bootstrap score is set by default to 60% on the Treeview application, which could have included some still-related, but not as highly related branches compared to the 70% bootstrap score done in previous labs.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
Autophagy-related protein 101 [Arabidopsis arenosa]	Arabidopsis arenosa	408	408	80%	4.00E-134	99.51	215	KAG7607589.1
unnamed protein product [Arabidopsis arenosa]	Arabidopsis arenosa	400	400	80%	4.00E-131	98.05	216	CAE6263256.1
PREDICTED: autophagy-related protein 101 [Camelina sativa]	Camelina sativa	398	398	80%	3.00E-130	97.56	216	XP_010464727.1
PREDICTED: autophagy-related protein 101 isoform X1 [Camelina sativa]	Camelina sativa	398	398	80%	3.00E-130	97.07	216	XP_010444709.1
autophagy-related protein 101 [Capsella rubella]	Capsella rubella	396	396	80%	2.00E-129	97.07	216	XP_006281085.1
PREDICTED: autophagy-related protein 101 isoform X2 [Camelina sativa]	Camelina sativa	345	345	71%	3.00E-110	96.69	192	XP_010444710.1
autophagy-related protein 101 [Eutrema salsugineum]	Eutrema salsugineum	384	384	80%	9.00E-125	94.15	216	XP_024007163.1
hypothetical protein EUTSA_v10005270mg [Eutrema salsugineum]	Eutrema salsugineum	385	385	81%	6.00E-125	93.69	218	ESQ31178.1
hypothetical protein N665_0205s0056 [Sinapis alba]	Sinapis alba	354	354	80%	2.00E-113	91.75	217	KAF8101463.1
unnamed protein product [Microthlaspi erraticum]	Microthlaspi erraticum	377	377	80%	5.00E-122	91.71	216	CAA7056898.1
hypothetical protein Bca52824_020085 [Brassica carinata]	Brassica carinata	351	351	80%	6.00E-112	91.26	217	KAG2316963.1
unnamed protein product [Arabis nemorensis]	Arabis nemorensis	367	367	80%	3.00E-118	91.22	219	VVB18257.1
Autophagy-related protein 101 [Hirschfeldia incana]	Hirschfeldia incana	348	348	80%	5.00E-111	90.78	217	KAJ0236562.1
autophagy-related protein 101 [Raphanus sativus]	Raphanus sativus	342	342	80%	8.00E-109	90.34	218	XP_018454481.1
hypothetical protein BRARA_I00874 [Brassica rapa]	Brassica rapa	364	364	80%	6.00E-117	90.24	217	RID44053.1

(Figure 1: Table of orthologous sequences and their associated BLAST statistical values. Top 15 most relevant orthologus sequences were determined by e-value. The lower the e-value the higher confidence it is an ortholog match. The Query cover is the amount the specific sequence covers in relation to the query. The e-value is the amount of confidence we have that the sequence is related. The percent identity is the percentage related to the query. The acc. Length is the amino acid length of the sequences.).



(Figure 2: Multiple Sequence Alignment of top 15 orthologs from BLASTp search of accession number At5g66930, including only one other *Arabidopsis thaliana* sequence not the same as the query. [Most large conserved stranded regions were trimmed for visibility purposes. Colouring is the BLOSUM80 scheme, with a gradient legend; blue representing a better match and green representing a worse match. Dots represent conserved regions. Query sequence included on top row. Columns are row 47 to 251 for non-redundancy purposes.).


Daphnia pulex (6669.DappuP312785)
Branchiostoma floridae (7739.JGI126010)
Branchiostoma floridae (7739.JGI126021)
Nematostella vectensis (45351.NEMVEDRAFT_v1g217973-PA)
Hydra magnipapillata (6085.XP_002167371)
Trichoplax adhaerens (10228.TriadP54105)
Strongylocentrotus purpuratus (7668.SPU_016633tr)
Strongylocentrotus purpuratus (7668.SPU_013365tr)
Ciona savignyi (51511.ENSCSAVP00000011400)
Ciona intestinalis (7719.ENSCINP00000035803)
Petromyzon marinus (7757.ENSPMAP00000006833)
Danio rerio (7955.ENSDARP00000103018)
Gadus morhua (8049.ENSGMOP00000003903)
Oreochromis niloticus (8128.ENSONIP00000009923)
Gasterosteus aculeatus (69293.ENSGACP00000005927)
Tetraodon nigroviridis (99883.ENSTNIP00000008530)
Takifugu rubripes (31033.ENSTRUP00000029741)
Xiphophorus maculatus (8083.ENSXMAP00000007211)
Oryzias latipes (8090.ENSORLP00000021771)
Latimeria chalumnae (7897.ENSLACP00000021045)
Xenopus (Silurana) tropicalis (8364.ENSXETP00000053091)
Ornithorhynchus anatinus (9258.ENSOANP00000015194)
Monodelphis domestica (13616.ENSMODP00000032296)
Macropus eugenii (9315.ENSMEUP00000005072)
Sarcophilus harrisii (9305.ENSSHAP00000018667)
Dasypus novemcinctus (9361.ENSDNOP00000011178)
Echinops telfairi (9371.ENSETEP00000005953)
Procavia capensis (9813.ENSPCAP00000009886)
Loxodonta africana (9785.ENSLAFP00000023898)
Microcebus murinus (30608.ENSMICP00000006810)
Otolemur garnettii (30611.ENSOGAP00000013460)
Tarsius syrichta (9478.ENSTSYP00000011163)
Callithrix jacchus (9483.ENSCJAP00000037059)
Macaca mulatta (9544.ENSMMUP00000001688)
Nomascus leucogenys (61853.ENSNLEP00000001585)
Pongo abelii (9601.ENSPPYP00000009641)
Gorilla gorilla (9593.ENSGGOP00000014253)
Homo sapiens (9606.ENSP00000299886)
Pan troglodytes (9598.ENSPTRP00000016297)
Cavia porcellus (10141.ENSCPOP00000018381)
Cavia porcellus (10141.ENSCPOP00000003239)
Mus musculus (10090.ENSMUSP00000018805)
Rattus norvegicus (10116.ENSRNOP00000003804)
Ictidomys tridecemlineatus (43179.ENSSTOP00000001636)
Dipodomys ordii (10020.ENSDORP00000002346)
Oryctolagus cuniculus (9986.ENSOCUP00000010215)
Tupaia belangeri (37347.ENSTBEP00000010812)
Sus scrofa (9823.ENSSSCP00000018275)
Tursiops truncatus (9739.ENSTTRP00000004916)
Bos taurus (9913.ENSBTAP00000007999)
Equus caballus (9796.ENSECAP00000006507)
Felis catus (9685.ENSFCAP00000004377)
Canis lupus familiaris (9615.ENSCAFP00000006718)
Ailuropoda melanoleuca (9646.ENSAMEP00000002816)
Mustela putorius furo (9669.ENSMPUP00000014182)
Myotis lucifugus (59463.ENSMLUP00000015199)
Pteropus vampyrus (132908.ENSPVAP00000006353)
Anolis carolinensis (28377.ENSACAP00000014302)
Pelodiscus sinensis (13735.ENSPSIP00000020553)
Taeniopygia guttata (59729.ENSTGUP00000002997)
Meleagris gallopavo (9103.ENSMGAP00000006649)
Gallus gallus (9031.ENSGALP00000007041)
Tribolium castaneum (7070.TC009561-PA)
Acyrthosiphon pisum (7029.ACYPI001869-PA)
Heliconius melpomene (34740.HMEL012959-PA)
Danaus plexippus (13037.EHJ66433)
Bombyx mori (7091.BGIBMGA012450-TA)
Nasonia vitripennis (7425.NV10250-PA)
Apis mellifera (7460.GB18353-PA)
Atta cephalotes (12957.ACEP_00009457-PA)
Pediculus humanus (121225.PHUM413450-PA)
Anopheles darlingi (43151.ADAR004533-PA)
Anopheles gambiae (7165.AGAP001322-PA)
Culex quinquefasciatus (7176.CPIJ007948-PA)
Aedes aegypti (7159.AAEL007141-PA)
Drosophila willistoni (7260.FBpp0240708)
Drosophila grimshawi (7222.FBpp0149372)
Drosophila virilis (7244.FBpp0224481)
Drosophila pseudoobscura (7237.FBpp0281462)
Drosophila ananassae (7217.FBpp0120932)
Drosophila yakuba (7245.FBpp0269552)
Drosophila melanogaster (7227.FBpp0082093)
0.68

(Page above: Figure 3: Phylogenetic tree of accession number At5g66930, bootstrap values of 0.60, or 60%, scale bar included at the bottom of 0.68 provided by Treeview.)

In this phylogenetic tree, a pertinent trend is that the nodes root out into 2 taxons throughout, as well as many events of parallel evolution, where there is independent evolution of the same character from a different ancestral state. It can also be noted that the bootstrap values are generally higher; around 0.90 near the terminal tips compared to the based of the tree.

Discussion:

The unknown function of this gene is highly likely to be related to one of the functions of the closely related orthologs listed in the 3 figures above. Investigating the function of the lowest e-value orthologs, we see that the vast majority of them have autophagy involved or related to it. Autophagy is the natural process in which the cell breaks down unusable proteins in the cytoplasm. Therefore, the unknown process could possibly be autophagy, with the function being protein binding at a molecular level. The relationships with other taxa make great sense to me, as the orthologs that are closely related, indicated by the high bootstrap values in the phylogenetic tree in addition to highly conserved reading frames. All the taxa involve autophagy and the function is essential for life. The At5g66930 gene is widespread in the tree of life, especially for plants because autophagy allows for cellular health maintenance. Thus, we can infer that the gene is relatively old, because it has been involved with this mechanism for species that have been around for a long time. The gene is highly conserved across taxa, as seen in the BLAST homology search (Fig. 2) as well as in the practical sense, it is a mechanism that must be conserved for species to maintain cellular health. Some regions are highly conserved, with

regions of multiple sites in a row being conserved, with others having some columns that are mutated; as seen in (Fig. 2). This can be interpreted in a way such that since it is highly conserved for the most part of the orthologous sequences, the function is greatly important, if not essential for the related taxa. By observing the NCBI Blast search results[2], we can see that there are paralogs of the provided gene, which indicates a duplication event. In addition, there are some matches after the BLAST that result in paralogs; referring back to the one *Arabidopsis thaliana* paralog included in Figure 1 and Figure 2. This means that these species have duplicatged versions of this gene. The duplications occur *sometimes* whenever the phylogenetic tree (Figure 3) branches off n times from a node. It is only sometimes because some events are speciation events, indicating orthologs; not paralogs.

References

1. Jaime Huerta-Cepas, Francois Serra and Peer Bork. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. Mol Biol Evol 2016; doi: 10.1093/molbev/msw046 (2023).
2. National Center for Biotechnology Information (NCBI): National Library of Medicine (US), National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/ (1988).
3. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang,
   Webb Miller, and David J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402 (1997).
4. Tamura K, Stecher G, and Kumar, MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution* 38:3022-3027 (2021).
5. The Arabidopsis Information Resource (TAIR), https://www.arabidopsis.org/servlets/TairObject?type=locus&name=AT5G66930, on www.arabidopsis.org, (2023).