S1803764 October 2021

Bioinformatics 1 – coursework 1
Report

Introduction (250 word limit) [10 marks]

L-gulonolactone oxidase is an enzyme which synthesizes ascorbic acid (vitamin C) from glucose. Ascorbic acid is necessary for the production of hydroxysline and hydroxyproline which are both essential for collagen synthesis in connective tissues. Abnormal collagen results in losing teeth, leaking blood vessels and hemorrhage in the joints, gums and intestines.[9]

Vitamin C is considered an essential vitamin for all life on Earth. Nearly all animals on the planet make plenty of their own vitamin C and thus have no need for it in their diets. This is unlike humans, primates, and a select few other animals who need to obtain vitamin C in their diet.

Although humans still have the genes necessary for vitamin C biosynthesis, one of them (the GULO gene) was randomly mutated beyond repair resulting in the gene being no longer functional, rendering it a pseudogene.

These random mutations happen regularly and are a natural part of evolution, as the strength of the mutation typically corresponds to the health/strength of the individual (strong genes remain and weak ones die off). Thus it is very surprising that this GULO mutation has survived across multiple different species, the only reasonable explanation for such a phenomenon is that the species that lost this gene were species that already had a plentiful amount of vitamin C in their diet, and thus this gene was not essential for their survival.

My research in this report is trying investigate whether a select set of species have or do not have a functioning GULO gene as a means to learn more about their evolution.

Materials & Methods

When starting this research I first decided to get some more technical background information on the GULO gene in *mus musculus*, and it's homology to all the other species we are looking to compare with: *homo sapiens*, *xenopus tropicalis*, *elephas maximus*, *cavia porcellus*, and *branchiostoma lanceolatum.*

This background information is extremely important as it helps us determine what BLAST methods and parameters would be most effective and reliable for sequence comparison for each of these species. This will not only make our BLAST searches far more efficient but also give us a far greater insight into the results we get.

The first piece of information I looked to obtain was the protein coding sequence for this GULO gene. This is due to the fact that protein sequences prove far more useful than nucleotide sequences when researching the homology between different species. This can be attributed to a number of factors that differentiate protein and DNA sequences: 20 amino acids instead of 4 bases and a shorter sequence input increase the sensitivity of your results significantly for protein sequences, the degeneracy or redundancy of codons may lead to "noise" in DNA sequences, by using a protein weight matrix we can also take physico-chemical properties into consideration (some amino acid substitutions are more likely to occur than others), and DNA sequences may include non-coding regions that can lead to false results.[7] I obtained this protein coding sequence by performing a high-similarity BLASTN search with the provided DNA sequence (XM_006519066.1) as input, the first hit was evidently the correct *mus musculus* GULO gene as it had 100%

identity and 100% query coverage, given this I went on the NCBI page for this gene and retrieved it's protein accession number (NP_848862.1) in the 'General protein information' section.

Next I decided to investigate what BLAST methods and parameters should be most effective for researching homology between all these different species. For BLASTN I decided to always initially use megablast as my program selection as this focuses on retrieving highly similar sequences, if this program selection produced no hits I would then try discontiguous megablast and traditional blastn as these are more intended for cross-species comparisons (where the 3rd base wobble position is taken into account by focusing on finding matches at the first and second codon positions whilst ignoring the mismatches in the third position). For BLASTP I decided to always try blastp for program selection initially and then try psi-blast and delta-blast as these are more sensitive algorithms which can be used to find distant relatives in sequences. All the BLAST algorithm uses "words" as a means to nucleate regions of similarity in sequences, so throughout all BLASTP and TBLASTN searches I always set the word size to be 3 as this would maximise hits, and in BLASTN searches I used all default word sizes provided by the chosen program selection. Lastly, throughout all BLAST searches I always initially set the expectation threshold to be 0.05 as this ensures only significant hits with high confidence are found. If no hits were found at this threshold value I still attempted another BLAST at threshold 10 to ensure I found at least 1 hit for the given organism as this could just be a result of high evolutionary distance and although hits at such a threshold are unlikely to be noteworthy there is still a chance that this alignment is a result of a homologous relationship.

Finally, we need to be able understand the output metrics given by BLAST to make reliable deductions from them. The main metric I will use to determine the best BLAST hits is the e-value as this provides the likelihood that a given sequence match is purely by chance, this is extremely useful as it indicates how confident we can be about whether a sequence match is due to homology. I will use this in conjunction with bit score, query coverage % and identity %. The bit score is also a very useful metric as it measures sequence similarity independent of query sequence length and database size and is normalised based on the raw pairwise alignment score, this allows us to reliably compare bit scores of BLAST hits when using different BLAST parameters. Lastly, query coverage % and identity % are also extremely useful metrics when used together as they indicate how much of the input sequence we managed to align with (query coverage) and how many characters within the covered part of the input sequence were identical (identity %), and thus give a great insight into the true quality of the alignment.

Results

Homo Sapiens

Due to *homo sapiens*' inability to synthesise their own vitamin C it is easy to assume that *homo sapiens* lack a functioning GULO gene. However, I do believe we will still get some significant sequence matches in BLAST against the *mus musculus*' GULO gene as remnants of this gene is still present in *homo sapiens* it has just been mutated to the point that it has lost it's function (rendering it a pseudogene).

BLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
XM_006519066.1	nr	0.05	High similarity	0	-	-	-	-	-	-	-
XM_006519066.1	nr	0.05	Somewhat Similarity	12	189	529	40%	85.98%	2e-44	NG_001136.2	GULOP gulonolactone (L-) oxidase,

						pseudogene
189	703	53%	85.98%	2e-44	AF311103.5	Homo sapiens chromosome 8 clone Scb-212E3 map p21-p11, complete sequence
189	434	32%	85.98%	2e-44	D17461.1	Homo sapiens GULO gene for L-gulono-gamma-lactone oxidase, exons 9, 10 and 12

Table 1.1: In this nucleotide blast I found a hit on the human's GULO pseudogene sequence NG_001136.2. Thus it is evident that the GULO gene is present in homo sapiens, however, it is not functional. This is evident given "pseudogene" is present in the name for this sequence and also from the quality of results we received. If this gene was still functional we would expect a far larger query coverage percentage, and identity percentage, as homo sapiens' GULO gene sequence would not have not have been mutated so much from mus musculus' GULO gene sequence. An interesting result from this BLAST search was the "chromosome 8 clone" sequence as this achieved the highest total bit score and query coverage across all the sequences, this implies that the human's GULO pseudogene is stored inside chromosome 8.

BLASTP												
Parameters					**# Hits**	**Best Hit(s)**						
Input	**Database**	**Expect Threshold**	**Scoring Matrix**	**Program Selection**		**Max Bit Score**	**Total Bit Score**	**Query Coverage**	**Percentage Identity**	**E Value**	**Sequence ID**	**Sequence Name**
NP_848862.1	nr	0.05	BLOSUM62	Psi-blast	46	67	67	26%	31.09%	9e-11	NP_055577.1	delta(24)-sterol reductase precursor
NP_848862.1	nr	0.05	BLOSUM62	Delta-blast	64	229	229	94%	15.61%	6e-69	NP_919417.1	Probable D-lactate dehydrogenase, mitochondrial isoform 2 precursor

Table 1.2: In this protein blast there was no hit against a GULO ortholog thus we can expect this is due to the fact that the GULO DNA sequence in *homo sapiens* has been mutated beyond function to the point that it is unable to be used for translation into a GULO protein. I also found an interesting hit with the LDH enzyme sequence, this seems rather unexpected given it obtained 94% query coverage (yet only 15.61% identity) and an e value of 6e-69 despite it not being a GULO gene isoform. After conducting some research it seems that there is some relationship between the GULO gene and this enzyme as medicines containing ascorbic acid (vitamin C) are proven to decrease your LDH levels.[6]

TBLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	5	68.6	210	32%	77.5%	1e-9	D17461.1	Homo sapiens GULO gene for L-gulono-gamma-lactone oxidase, exons 9,10 and 12
					68.6	261	40%	77.5%	1e-9	NG_001136.2	Homo sapiens gulonolactone (L-) oxidase, pseudogene (GULOP) on chromosome 8
					68.6	322	48%	77.5%	1e-9	AF311103.5	Homo sapiens chromosome 8 clone Scb-212E3 map p21-p11, complete sequence

Table 1.3: In the protein-nucleotide blast I found a hit with the human's GULO pseudogene sequence again which reaffirms the idea that although this GULO gene is still present in homo sapiens it is no longer functional.

Analysis of these results

It is evident that *homo sapiens* possess this GULO gene based on the BLAST matches found above. However, the gene found in these matches was referred to as a pseudogene indicating that this gene is non-functional, this is also suggested by the query coverage percentages found across our various BLAST searches, these query percentages are much lower than we would expect if this gene was still functional, as such low percentages suggest a significant number of mutations from the GULO gene possessed by *mus musculus* which has a very high chance of hindering it's function. On top of this we find very significant hits with *homo sapiens*' "chromosome 8 clone" sequence indicating that this pseudogene must be stored in *homo sapiens*' chromosome 8.

Xenopus tropicalis

The *xenopus tropicalis* is related to *mus musculus* by being in the *tetrapod* superclass, this is a very large group of species and thus we can expect some more significant differences in the GULO gene sequences between these species due to their large evolutionary distance. This is evident as amphibians (such as *xenopus tropicalis*) produce vitamin C in their kidneys, in contrast to mammals (such as *mus musculus*) who produce vitamin C in their liver.[8]

BLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
XM_006519066.1	nr	0.05	High similarity	0	-	-	-	-	-	-	-
XM_006519066.1	nr	0.05	Somewhat similarity	2	737	737	98%	72.86%	0	XM_031903103.1	PREDICTED: Xenopus tropicalis L-gulonolactone oxidase (LOC101733882), mRNA

Table 2.1: In the nucleotide blast when changing program selection to traditional blastn (somewhat similarity) I found a very strong hit with *xenopus tropicalis*' GULO gene. This hit had an extremely high bit score, query coverage and identity percentage, these metrics in conjunction with an e value of 0 means we can say with high confidence that this match is a result of a homologous relationship. One thing we should

note is that this sequence has been “predicted” as denoted by the title of this sequence meaning that this sequence is likely to have been a translation from *xenopus tropicalis’* GULO protein sequence meaning it has not actually been verified by experimentation.

BLASTP												
Parameters					# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	blastp	19	683	683	100%	71.82%	0	XP_031758963.1	L-gulonolactone oxidase

Table 2.2: In the protein blast I found an extremely good sequence match with 100% query coverage and almost 72% identity on *xenopus tropicalis*’ GULO protein. With an e value of 0 it is safe to assume this GULO gene has been preserved in this species and given the level of identity it implies there are very few mutations between this species’ GULO gene and that of the *mus musculus* and so it is highly likely this gene is still functional.

TBLASTN										
Parameters				# Hits	Best Hit(s)					
Input	Database	Expect Threshold	Scoring Matrix		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	10	658	658	100%	69.55%	0	XM_031903103.1	PREDICTED: Xenopus tropicalis L-gulonolactone oxidase (LOC101733882), mRNA

Table 2.3: In the protein-nucleotide blast I found an extremely good sequence match again with an e value of 0, 100% query coverage and almost 70% identity. Again we should note this sequence is a prediction and has not been verified by experimentation.

Analysis of these results

From all the BLAST searches above it is evident that *xenopus tropicalis* still possesses this GULO gene. Throughout all these searches the best hit had an average e value of 0, query coverage of 99%, and identity of 71%, however, the only result we can really trust in this instance is that retrieved from BLASTP. This is due to the fact that the other sequence hits in BLASTN and TBLASTN were computational predictions as to what *xenopus tropicalis*’ GULO nucleotide sequence looked like and were not actually verified by experimentation. However, even just using BLASTP’s result alone I think it is safe to assume that this GULO gene is still functional in *xenopus tropicalis* based on the extremely good bit score, query coverage, and identity values we received. We can expect this 29% mismatching identity to be attributed to the large evolutionary distance between *mus musculus* and *xenopus tropicalis*.

Elephas maximus

Elephas maximus is related to *mus musculus* by being in the *mammalia* class of verterbrates. These species are quite close together in evolutionary distance so we can expect some good sequence alignments if *elephas maximus* has the GULO gene. The *elephas maximus* lives in a freshwater habitat with a diet that includes fruit which may indicate that they not actually require this GULO gene as they obtain vitamin C in their diet.

BLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
XM_006519066.1	nr	10	High similarity	0	-	-	-		-	-	-
XM_006519066.1	nr	10	Somewhat similarity	3	29.2	29.2	1%	89.96%	2.8	JX941470.1	Elephas maximus borneensis microsatellite Em_0925

Table 3.1: In the nucleotide blast I only found a hit after setting the expectation threshold to 10, resulting in 3 hits of seemingly unrelated genes to the GULO given the best one only achieved 1% query coverage. An e value of 2.8 suggests that this match is not significant and is not necessarily a result of homology.

BLASTP												
Parameters					# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	blastp	0	-	-	-	-	-	-	-
NP_848862.1	nr	10	BLOSUM62	blastp	18	23.1	23.1	4%	44.44%	1.6	AAG17885.1	NADH dehydrogenase subunit 5, partial (mitochondrion)

Table 3.2: In the protein blast again I only found a hit after setting the expectation threshold to 10. The best hit from this BLAST search only gave a query coverage of 4% and was not the GULO gene so I think it is safe to assume that *elephas maximus* lacks the GULO gene altogether.

TBLASTN										
Parameters				# Hits	Best Hit(s)					
Input	Database	Expect Threshold	Scoring Matrix		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	0	-	-	-	-	-	-	-
NP_848862.1	nr	10	BLOSUM62	4	24.3	24.3	8%	27.03%	0.67	GU557845.1	Elephas maximus voucher Ema-2 clone 316 locys B256 genomic sequence

Table 3.3: As before, in the protein-nucleotide blast I only found a hit after increasing the threshold to 10, and the best hit here was still not a match against *mus musculus*' GULO gene (as given by the sequence name). This sequence match had an e value of 0.67 indicating that this match is not necessarily significant but may be a result of homology, which is to be expected given the relatively small evolutionary distance between *mus musculus* and *elephas maximus*.

Analysis of these results

It is evident that *elephas maximus* does not possess this GULO gene given we had no hits with any significant similarity throughout all our BLAST searches. Given the relatively small evolutionary distance between *mus musculus* and *elephas maximus* this seems rather surprising as I would have at the least expected some GULO pseudogene to be preserved from when this gene was mutated and lost in this species. I believe this result can partly be attributed to the relative amounts of research and data available for these

species. After conducting some research on this species on NCBI I came to find there was a very large disparity in the amount of data between these species: on NCBI *mus musculus* has 11,526,868 nucleotides in comparison to *elephas maximus*' 4,933, and *mus musculus* has 1,404,013 proteins in comparison to *elephas maximus*' 14,962. From these results I think it is evident that this species is rather under documented and this may be the reason as to why we found such poor matches in our BLAST searches. It is likely this GULO gene still exists in *elephas maximus* (likely as a pseudogene based on *elephas maximus*' vitamin C rich diet) but it has just not been recorded on NCBI. Thus given the background of *elephas maximus*, relation to mus musculus, and results of our BLAST searches, I think it is best to say the results for *elephas maximus* are inconclusive.

Cavia porcellus

I expect very significant sequence matches for this species purely based on its close homology with *mus musculus* as they are both mammals of the order *rodentia*. Given they are part of the same taxon it is far more likely this GULO gene was preserved in this species as it implies a smaller evolutionary distance than other species listed here and thus less gene mutations.

BLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
XM_006519066.1	nr	0.05	High similarity	2	743	743	54%	85.28%	0	XM_013143314.2	PREDICTED: Cavia porcellus L-gulonolactone oxidase-like (LOC100735706), mRNA
					213	399	23%	91.08%	9e-54	D12762.2	Cavia porcellus L-gulono-gamma-lactone oxidase gene homologue, exon 2, 3, 4, 5, 7, 8, 9 corresponding parts

Table 4.1: In this nucleotide blast I found a very significant hit with cavia porcellus' GULO gene sequence XM_013143314.2 which is evident given it produced an e value of 0. However, this hit only achieves 54% query coverage against the original GULO sequence, and such high variability is very unlikely to preserve the function of this gene. The second best hit is also some version of the GULO nucleotide sequence (that is not a result of a computational prediction), however, it also has an insufficient query coverage to be able to preserve the function in this gene. This is especially true when considering how alike these sequences should be if we assumed cavia porcellus still had a functional GULO gene given how close these species are in evolutionary distance (implying less genetic mutations).

BLASTP												
Parameters					# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	blastp	7	615	615	99%	65.65%	0	XP_012998768.1	LOW QUALITY PROTEIN: L-gulonolactone oxidase-like

Table 4.2: In the protein blast I found a significant hit on a GULO protein sequence with extremely strong query coverage and percentage identity, however, this protein has been labelled as "low quality". This label is a result of a computational prediction and was derived from a genomic sequence. On top of this we must note that 11% of the sequence alignment for this protein were gaps, meaning a very significant amount of

information was lost from *mus musculus*' GULO protein sequence.

TBLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	6	578	578	99%	62.6%	0	XM_013143314.2	PREDICTED: Cavia porcellus L-gulonolactone oxidase-like (LOC100735706), mRNA
					95.9	720	80%	84%	3e-20	D12762.2	Cavia porcellus L-gulono-gamma-lactone oxidase gene homologue, exon 2, 3, 4, 5, 7, 8, 9 corresponding parts

Table 4.3: In the protein-nucleotide blast I found a very powerful hit on a GULO protein sequence. However, one thing we must note is that this hit was on a computational prediction for a sequence and so was not actually verified by experimentation, the sequence used for this translation was the low quality protein we found in the protein blast (table 4.2).

Analysis of these results

It is evident that *cavia porcellus* does not possess this GULO gene. Although we got some very strong hits on a GULO sequence, these hits were with a GULO "like" sequence that is still present in cavia porcellus as described in the profile for this gene on NCBI. This is further reinforced by the "LOW QUALITY PROTEIN" labelling from blastp and the high gap percentages we received for this sequence alignment. If the functionality of this GULO gene were to be conserved in *cavia porcellus* we would expect 100% query coverage and very few gaps. This is especially for species that are so closely related as this would imply that were only minimal/synonymous mutations from the original sequence.

Branchiostoma lanceolatum

The *branchiostoma lanceolatum* is related to *mus musculus* by being in the *Chordata* phylum, this is the phylum of the animal kingdom and so evidently includes a very large number of animal species. This indicates there is a very large evolutionary distance between these two species implying it is likely that there is high variability between the sequences for this gene in these species.

BLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
XM_006519066.1	nr	10	High Similarity	0	-	-	-		-	-	-
XM_006519066.1	nr	10	Somewhat Similarity	2	30.1	30.1	1%	90.48%	0.98	FM864151.1	Branchiostoma lanceolatum partial GpHR gene for glycoprotein hormone receptor-related protein

Table 5.1: In the nucleotide blast I only received hits after setting the expectation threshold to 10. All the hits were thus very poor and unrelated to the GULO gene.

BLASTP												
Parameters					# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix	Program Selection		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	10	BLOSUM62	blastp	3	25	25	10%	32.65%	1.3	O02466.1	Insulin-like peptide receptor
NP_848862.1	nr	10	BLOSUM62	Delta-blast	3	25.4	25.4	12%	15.09%	0.98	AAM93476.1	H+-transporting ATP synthase alpha subunit isoform 1, partial
NP_848862.1	refseq_protein	10	PAM250	blastp	3	22.7	22.7	4%	35%	0.12	NP_0075450.1	Cytochrome c oxidase subunit II (mitochondrion)

Table 5.2: In the protein blast I only received hits after setting the expectation threshold to 10, and there were no GULO orthologs present in these.

TBLASTN											
Parameters				# Hits	Best Hit(s)						
Input	Database	Expect Threshold	Scoring Matrix		Max Bit Score	Total Bit Score	Query Coverage	Percentage Identity	E Value	Sequence ID	Sequence Name
NP_848862.1	nr	0.05	BLOSUM62	0	-	-	-	-	-	-	-
NP_848862.1	nr	10	BLOSUM62	6	26.6	26.6	12%	32.76%	0.95	JQ942473.1	Branchiostoma lanceolatum GATA binding protein 1/2/3 mRNA, partial cds

Table 5.3: In the protein-nucleotide blast I only received hits after setting the expectation threshold to 10, and there were no GULO orthologs present in these.

Analysis of these results

I believe the results for this species are inconclusive due to the nature of this sequence comparison. The *branchiostoma lanceolatum* is an inverterbrate and thus it is far more difficult to compare with *mus musculus* than the other species as they are all vertebrates, such a characteristic can make a significant difference in the make up of the species. This is due to the fact that these animals were not nearly related at all in evolution and thus they may not have any significant similarities between their gene encoding. On top of this the *branchiostoma lanceolatum* species is very under documented in comparison to mus musculus as given by the statistics on NCBI where: *mus musculus* has 11,526,868 nucleotides in comparison to *branchiostoma lanceolatum*'s 60,290, and *mus musculus* has 1,404,013 proteins in comparison to *elephas maximus*' 697. From these results I think it is evident that this species is rather under documented and this may be the reason as to why we found such poor matches in our BLAST searches. It is likely this GULO gene still exists in *branchiostoma lanceolatum* (likely as a functional gene based on the lack of vitamin C in *branchiostoma lanceolatum*'s diet) but it has just not been recorded on NCBI.

Discussion (400 word limit) [20 marks]

From my research I found 1 species with a functional GULO gene (*xenopus tropicalis*), 2 species with GULO pseudogenes (*homo sapiens* and *cavia porcellus*), and 2 species whose results were inconclusive (*elephas maximus* and *branchiostoma lanceolatum*).

The findings for each species was based off a number of key factors: the quality of their blast hits, their evolutionary distance from *mus musculus*, and whether they have sufficient representation on NCBI. Evidently if any sequence returns incredible BLAST results with 100% query coverage and 0% gaps it is evident that this sequence is a match with *mus musculus*' GULO gene and must still be functional given such high similarity. However, if we do not get any good matches we must be careful as to why this may be as this could not be due the fact that this species lacks this GULO gene entirely but rather that this species has a very large evolutionary distance from *mus musculus*, or this species' genome is not fully documented on NCBI.

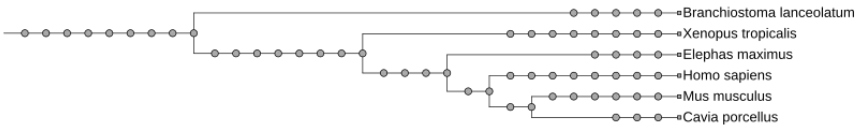

Figure 1: A phylogenetic tree to visualize the evolutionary distance between these species. Each circle on the branches of these clusters represents the various levels of taxonomic classification for each species.

As discussed above, I was easily able to tell that the functionality of the GULO gene in *xenopus tropicalis* was preserved purely based on the quality of BLAST hits I found which achieved 100% coverage and 0% gaps.

I was able to tell that the functionality of the GULO gene in *homo sapiens* and *cavia porcellus* were not preserved based on the quality of BLAST hits we found on their GULO genes given their relatively small evolutionary distance to *mus musculus*, and good genome documentation.

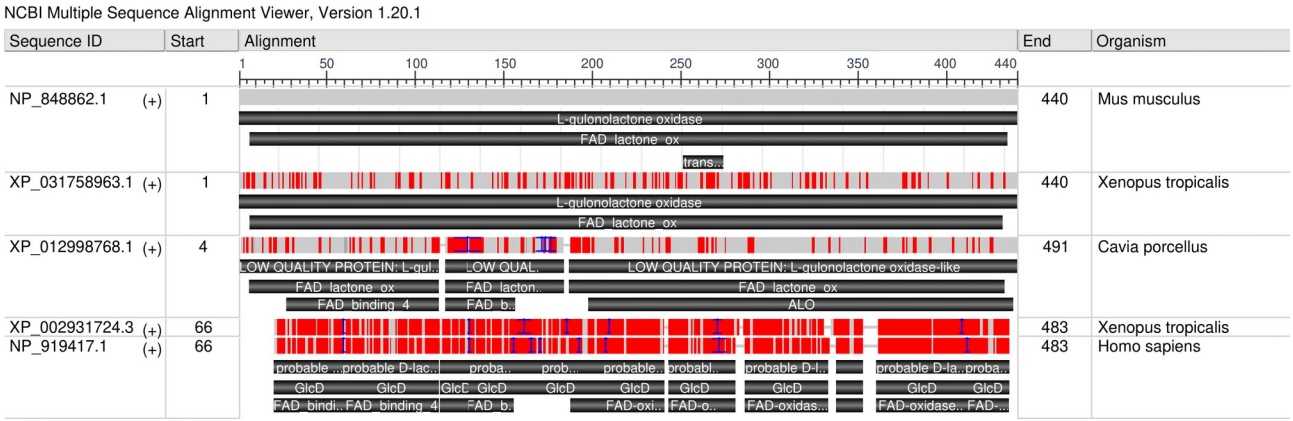

Figure 2: A multiple sequence alignment viewer to visualize the protein sequence alignments for *xenopus tropicalis*, *cavia porcellus*, and *homo sapiens*, with *mus musculus*' GULO protein sequence. The grey bars represent matches, red bars represent mismatches, grey dashes represent a gap, and blue bracket represents an insertion.

I decided to classify the findings of both *elephas maximus* and *branchiostoma lanceolatum* as inconclusive based on the fact I did not find any GULO orthologs in these organisms, this is rather surprising given how integral this gene is for survival, and thus we would at the least expect a match with a pseudogene. After further inspection I came to realise how under documented these organisms were on NCBI which would directly affect our results as without a recording of the complete genome for each of these species we cannot be sure whether this gene is or is not present in these organisms still. On top of this, we can also attribute the poor quality results achieved by *branchiostoma lanceolatum* to be due to it's large evolutionary distance

from *mus musculus* given that it is an invertebrate. This is evident as the sequence identity among GULO proteins for vertebrates varies from 64-95% in comparison to 47-95% for chordates, and unlike invertebrates vertebrate GULOs are also highly conserved in their exon-intron structures where all identified vertebrate GULO orthologs share 11 conserved exons.[1]

To improve upon my research in future I would aim to obtain the full genomes for *elephas maximus* and *branchiostoma lanceolatum* as a means to try find any existing GULO orthologs. I would also like to perform BLAST searches using GULO sequences from multiple species (not only mus musculus) as a means to get more reliable findings and get a better insight into the sequence similarities between different species in varying phylum.

References [5 marks]

1. Yang H. Conserved or lost: molecular evolution of the key gene GULO in vertebrate vitamin C biosynthesis. *Biochem Genet*. 2013;51(5-6):413-425. doi:10.1007/s10528-013-9574-0

2. Inai Y, Ohta Y, Nishikimi M. The whole structure of the human nonfunctional L-gulono-gamma-lactone oxidase gene--the gene responsible for scurvy--and the evolution of repetitive sequences thereon. *J Nutr Sci Vitaminol (Tokyo)*. 2003;49(5):315-319. doi:10.3177/jnsv.49.315

3. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *J Biol Chem*. 1994;269(18):13685-13688.

4. Nishikimi M, Koshizaka T, Ozawa T, Yagi K. Occurrence in humans and guinea pigs of the gene related to their missing enzyme L-gulono-gamma-lactone oxidase. *Arch Biochem Biophys*. 1988;267(2):842-846. doi:10.1016/0003-9861(88)90093-8

5. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol*. 2010;11(3):R26. doi:10.1186/gb-2010-11-3-r26

6. Lactate Dehydrogenase Isoenzymes. Ucihealth.org. http://www.ucihealth.org/health-library/content?contentTypeID=167&contentID=lactate_dehydrogenase_isoenzymes#:~:text=Medicines%20that%20contain%20ascorbic%20acid,can%20affect%20your%20LDH%20levels. Published 2021. Accessed October 25, 2021.

7. Brandt, Christian. (2016). Re: Why we always take amino acid sequence instead of nucleotide sequence for homology of two closely related species?. Retrieved from: https://www.researchgate.net/post/why_we_always_take_amino_acid_sequence_instead_of_nucleotide_sequence_for_homology_of_two_closely_related_species/56937cf161432547538b4567/citation/download.

8. Xie Y, Liu Y, Zhao Y et al. Gulo Acts as a de novo Marker for Pronephric Tubules in Xenopus laevis. *Kidney and Blood Pressure Research*. 2016;41(6):794-801. doi:10.1159/000450561

9. O'Malley B. Chapter 9 - Guinea Pigs. In: *Clinical Anatomy and Physiology of Exotic Species: Structure and Function of Mammals, Birds, Reptiles, and Amphibians*. Edinburgh u.a: Elsevier Saunders; 2007.

10.

Appendix

- “Sequence R” in the “input” table column in the “Results” section corresponds to the provided mus musculus’ DNA sequence:

>Sequence_R
ATGGTCCATGGGTACAAAGGGGTCCAGTTCCAAAACTGGGCGAAGACCTATGGCTGCAGTCCAGAGATGT
ACTACCAGCCCACATCAGTGGGGGAGGTCAGAGAGGTGCTGGCCCTGGCCCGGCAGCAGAACAAGAAAGT
GAAGGTGGTGGGTGGCGGCCACTCGCCTTCAGACATCGCCTGCACCGATGGCTTCATGATTCACATGGGC
AAGATGAACCGGGTTCTCCAGGTGGACAAGGAGAAGAAGCAGGTCACAGTGGAAGCCGGTATCCTCCTGA
CTGACCTGCACCCACAGCTGGACAAGCATGGCCTGGCCCTGTCTAATCTGGGAGCCGTGTCTGATGTGAC
GGTTGGTGGCGTCATTGGGTCTGGAACACATAACACCGGGATCAAGCACGGTATCCTGGCCACCCAGGTG
GTGGCCCTGACCCTGATGAAGGCTGATGGAACAGTTCTGGAATGTTCTGAGTCAAGTAATGCAGATGTGT
TCCAGGCTGCAAGGGTGCACCTGGGCTGCCTGGGTGTTATCCTCACTGTCACCCTGCAGTGTGTGCCACA
GTTCCACCTTCTGGAGACATCCTTTCCTTCGACCCTCAAGGAGGTCCTTGACAACCTGGACAGCCACCTG
AAGAAGTCTGAGTACTTCCGCTTCCTCTGGTTTCCTCACAGTGAGAACGTCAGCATCATCTACCAAGATC
ACACCAACAAGGAGCCCTCCTCTGCATCTAACTGGTTTTGGGACTATGCCATTGGGTTCTACCTCCTGGA
ATTCTTGCTCTGGACCAGCACCTACCTGCCACGCCTCGTGGGCTGGATCAACCGCTTCTTCTTCTGGCTG
CTGTTCAACTGCAAGAAGGAGAGCAGCAACCTCAGCCACAAGATCTTCTCCTACGAGTGTCGCTTCAAGC
AGCATGTCCAAGACTGGGCCATCCCCAGGGAGAAGACCAAGGAGGCCCTGCTGGAGCTAAAGGCCATGCT
GGAGGCCCACCCCAAGGTGGTAGCCCACTACCCCGTGGAGGTGCGCTTCACCCGAGGTGATGACATCCTG
CTGAGCCCGTGCTTCCAGAGGGACAGCTGCTACATGAACATCATTATGTACAGGCCCTATGGGAAGGATG
TGCCTCGGTTGGATTACTGGCTGGCCTATGAGACCATCATGAAGAAGTTTGGAGGCAGGCCCCACTGGGC
AAAGGCCCACAATTGCACCAGGAAGGACTTTGAGAAAATGTACCCCGCCTTTCACAAGTTCTGTGACATC
CGCGAGAAGCTGGACCCCACTGGAATGTTCTTGAATTCGTACCTGGAAAAGGTTTTCTACTAA

Formatting Guidelines

Please use a font size of 11 and normal single spacing. This works out to a report of maximum length of approximately 7-8 pages. Whilst we will not specify a minimum size, if you have produced a report of less than 5 pages it is unlikely to contain the necessary detail and discussion expected for the work. Please do not exceed this limit. In the report, you will need to carefully summarise and format your BLAST results for clarity and brevity (we will not accept pasted raw results from the BLAST web page in reports).

Reports can be prepared using whatever software you like Word, LibreOffice, LaTex etc. but please **only submit a single PDF format document for the report**.

Submission is made by clicking the link at the top of this section and uploading the PDF file. Please do not put your name on the file or the report, the LEARN system will handle your identity and enable anonymous marking.

If you have any questions please do ask in the discussion forum for coursework.