Bioinformatics approach for Investigating GULO functionality

B164439
INFR11160: Bioinformatics, University of Edinburgh

1 Introduction

Vitamin C is an important cofactor in many important physiological processes[1]. Vitamin C deficiency usually leads to diseases such as scurvy[2][3].The GULO gene encodes an enzyme which converts L-gulono-1,4-lactone to L-ascorbate (vitamin C)[4]. GULO enzyme (L-gulonolactone oxidase) is required in the terminal step of catalytic reaction[3][4]. Members of this enzyme family contain two important domains: FAD-binding domain and ALO domain[3]. FAD domain is critical for binding of flavin adenine dinucleotide (FAD) cofactor which is essential for transfer of electrons between L-gulono-γ-lactone (substrate)[5] during GULO enzyme catalysis whereas D-arabino-1,4-lactone oxidase (ALO) domain binds the substrate[1][4]. The molecular mechanism behind the GULO enzyme[5] is heavily studied as many mammals, including humans cannot synthesize vitamin C.

GULO enzyme shares same protein folding topology as members of vanillyl-alcohol oxidase (VAO) family which is known to adopt the lock and key mechanism in terms of substrate binding[6]. Therefore, when investigating the catalytic binding site of GULO enzyme, the structure of the enzyme must be of similar orientation in order for enzyme activity. Numerous publications have utilized phylogenetic analysis to prove GULO functionality in certain species[3]; Yang et. al, through phylogenetic analysis, suggested that inactivation of GULO gene through species was due to deleterious mutations of exons which disrupt GULO functionality[3]. Therefore, further investigation into the structure[7] and biochemistry of GULO enzyme using bioinformatics tools provides a better indication of how GULO functionality was maintained or lost. Here demonstrates a structural bioinformatics and functional genomics approach to determine the functionality of GULO gene in certain species.

2 Methodology

Detecting putative orthologs

BLAST algorithms (Blastn, Blastp and tBlastn) were implemented for query sequence provided (*Mus musculus* DNA coding sequence) against 4 different vertebrate species, *Homo sapiens*, *Xenopus tropicalis*, *Elephas maximus*, *Cavia porcellus* and one invertebrate, *Branchiostoma lanceolatum*. Blastx was used to query the DNA coding sequence against *Mus musculus* where the amino acid sequence of GULO enzyme was obtained (accession id: NP_848862.1). This was used as query sequence in the Blastp and tBlastn algorithms to gain putative GULO orthologs for 5 species mentioned.

The nr/nt databse was used in BLAST searches as it incorporates all sequences such as GenBank and RefSeq nucleotides[8]. Parameters which were modified to loosen/strengthen searches were BLOSUM matrices, match/mismatch scores, type of compositional adjustments and word size. BLOSUM matrices are chosen to minimize biases in databases as without clustering, some closely related sequences may be overrepresented[9]. BLOSUM also accounts for evolutionary deviation of varying time periods as compared to PAM matrices[8][9]. In figure 1 below, loosening parameters include decreasing match/ mismatch scores, BLOSUM matrices, gap costs and word size. This concept is adopted as match/ mismatch scores, and gap costs lowered indicates lenient penalties when constructing a matrix. Lowered word size indicates a less accurate search as probability of a shorter amino acid sequence existing in other species increases[10]. BLOSUM62 is lowered to BLOSUM45/50 as it yields results with only 45/50% identity[9]. For refinement purposes, composition-based statistics are used to improve the calculation of E-value statistics[11].

Inferring functionality from BLAST results

Results yielded from blastp, were consequently extracted to infer functionality. SMART database[12] was utilized to identify the functional domains of *Mus musculus*' L-gulono-γ-lactone oxidase which are FAD and ALO domains. Protein structures from Blastp results were extracted from UNIPROT: AlphaFold-predicted structures were used and proteins with no predicted structures will be predicted through the MODELLER server[13][14]. MODELLER is GUI-based bioinformatics tool used for structure prediction[13][15] which implements pairwise alignment of profile hidden Markov models (HMM)[15][16][14]. The structural bioinformatics approach involves looking at the catalytic sites of L-gulono-γ-lactone oxidase. PYMOL was used to visualize the structure of *Mus musculus* enzyme and the enzyme's catalytic sites. A comparison of GULO enzyme structure to other protein structures obtained from blastp would indicate if the GULO gene is functional in each species. Research groups have proposed methods of searching for catalytic

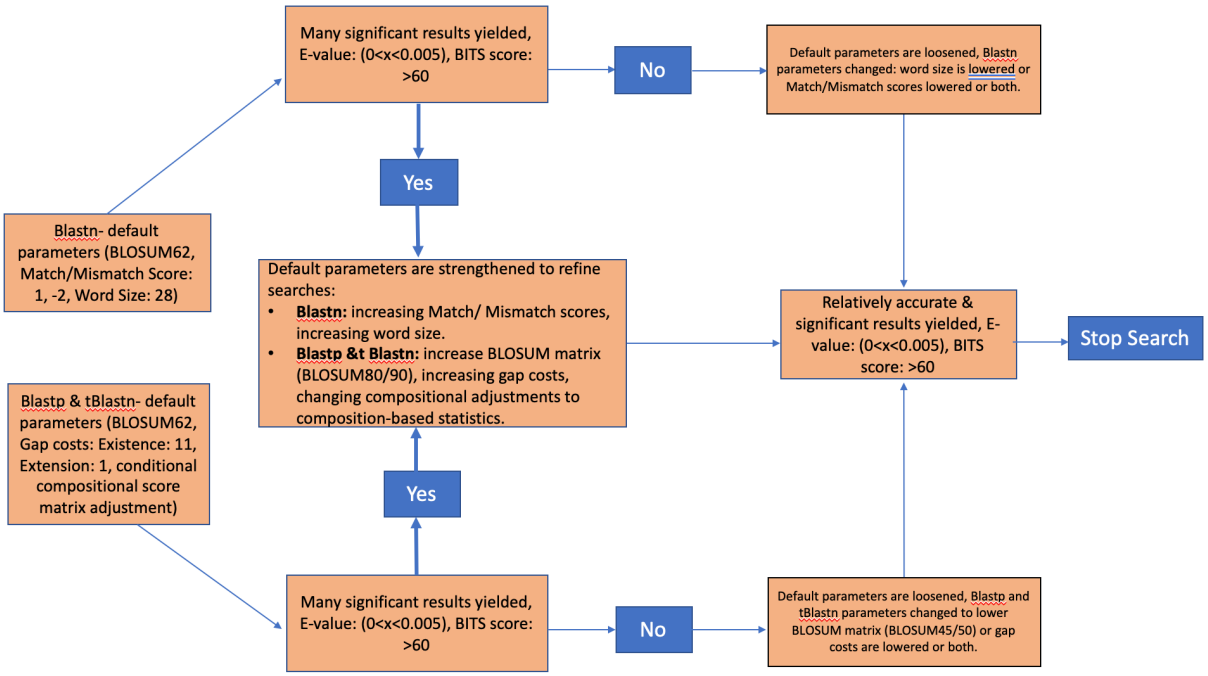

Figure 1: Iterative pipeline for optimizing BLAST parameters to gain accurate BLASTn searches against *Mus musculus* DNA coding sequence and amino acid sequence (obtained from Blastx) for Blastp and tBlastn searches.

motifs in enzymes as a way of determining function[17]. As phylogenetic analysis has shown there are two motifs important in the GULO enzyme in animals: HWAK[1][18] and VGGGHS[18] motif (where the letters indicate certain amino acids). Therefore, HWAK motif (part of ALO domain), VGGGHS motif (part of FAD domain) and both domains should be identified in the proteins extracted from Blastp results to infer functionality[7]. The catalytic site of this enzyme is at the interface of the two domains[5], buried in the core of the enzyme[1][5]. CONSURF analysis was carried out for *Mus musculus* amino acid sequence to identify the conserved regions of *Mus musculus*' GULO enzyme. ConSurf is a bioinformatics tool which reveals functional regions by analyzing evolutionary dynamics of amino acid substitutions in sequences[19].

3 Results

Mus musculus L-gulonolactone oxidase

Mus musculus amino acid sequence (accession id: AAH28822.1), investigated on UniPROT and SMART databases, showed two present domains responsible for the enzyme's activity as expected in this enzyme family. From figure 2A below, FAD and ALO domains are present from AlphaFold's predicted structure; VGGGHS and HWAK motifs are also present. When compared to figure 2B, ConSurf results indicate that these motifs are highly conserved in this region and at the interface of both domains, where enzyme functionality can be inferred[17][7] [19]. These highly conserved sites are presumed to be functionally or structurally important as they have undergone fewer mutations relative to rest of the alignment over evolutionary time[20]. This red region in figure 2B is the same as catalytic binding site containing both motifs in figure 1C, when structures are superimposed together. VGGGHS motif, annotated in Figure 2C, and 2D, is also proven to be critical in animals in order for L-ascorbic acid biosynthesis[18].

Moreover, figure 2E clearly supports published phylogenetic evidence that Histidine residue on FAD domain is completely conserved as H54 in *Mus musculus* amino acid sequence is required for covalent binding to FAD[3][18]. H54 has been reported to have FAD covalently linked to N(1) position of histidine when enzymatic reaction occurs[5][21]. Overall, from figure panel 2, it is indicative that if proteins which have been found in the 5 species below, have a similar protein structure, FAD and ALO domains, enzyme's active site at the interface of both domains, as well as the VGGGHS motif and HWAK motif, the GULO gene is functional.

3.1 *Homo sapiens*

BLAST searches for humans yielded significant BLAST results for 3 respective algorithms. Blastn results were obtained from the loosest parameters and the hit found had a low e-value but only covered 24% of the query. Pairwise alignment analysis shows multiple nucleotide substitutions and gaps

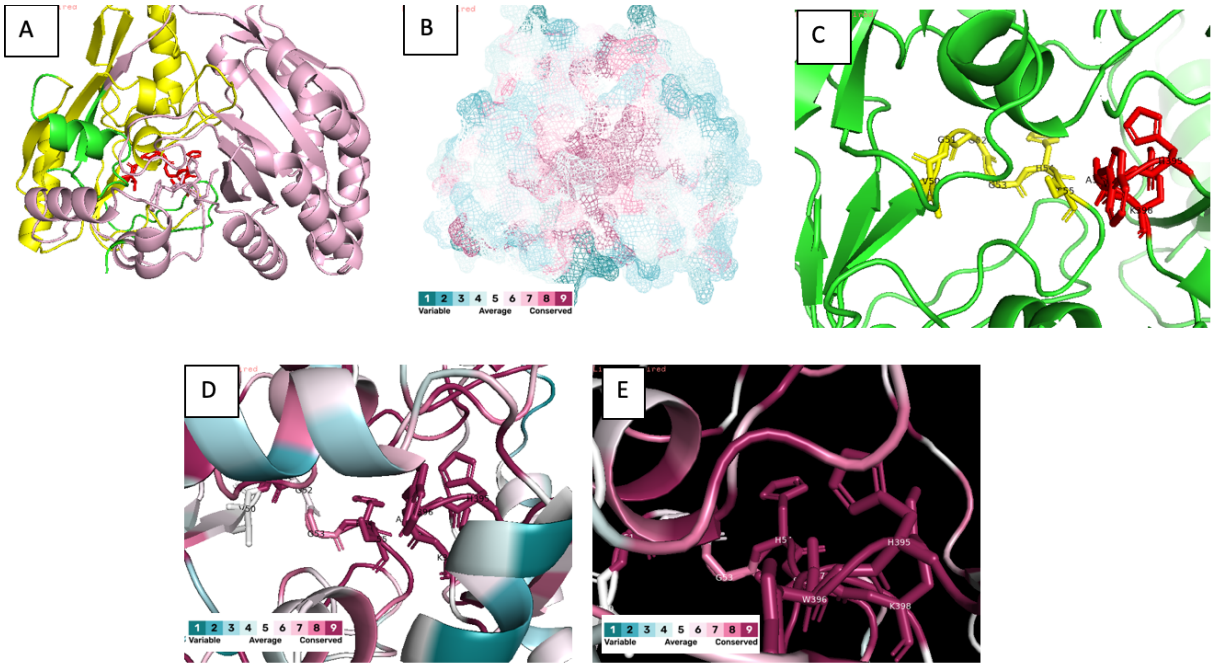

Figure 2: PYMOL-generated figures of *Mus musculus'* GULO enzyme: A) AlphaFold-predicted structure of *Mus musculus* L-gulonolactone oxidase, ALO domain (pink) and FAD domain (yellow), HWAK motif shown in pink sticks (part of ALO domain) and VGGGHS motif (FAD binding site) shown in red sticks, part of FAD domain. **B)** CONSURF analysis of *Mus musculus* L-gulonolactone oxidase, catalytic site is buried in the middle of the protein, as inferred from how conserved the sequences are **C)** Zoomed focus into enzyme's FAD binding site (VGGGHS motif shown in yellow sticks) and HWAK motif (red sticks) **D)** CONSURF analysis of FAD binding site indicating VGGGHS motif and HWAK motif are highly conserved and required for enzyme functionality **E)** CONSURF analysis and zoomed orientation (focus) on FAD binding residue, histidine (H54) which binds FAD.

which are deleterious as protein translation and folding from nucleotide sequence would be structurally and biochemically different to the GULO enzyme. Biomedical research papers have also provided a pairwise comparison with the *Mus musculus* sequence which have identified 2 single nucleotide deletions, 1 triple nucleotide deletion and one single nucleotide insertion in the human sequence found in *Homo sapiens* gulonolactone (L)-oxidase pseudogene (GULOP) on chromosome 8 (accession id: NG_001136.2)[22][23]. These indel mutations which have occurred may have induced a frameshift affecting the type of amino acid produced and consequently, the protein produced. This suggests that the human GULO gene has diverged more rapidly than species which possess a functional GULO. It would not be surprising if protein yielded from blastp search was structurally and functionally different from the GULO enzyme. When evaluating the literature found with the accession number, research has proven that this human sequence has deletion of a single amino acid, two stop codons and many amino acid substitutions[24][23][25].

A blastp search, using BLOSUM50 matrix, yielded delta(24)-sterol reductase precursor (accession id: NP_055577.1). However, this has low query coverage (28%) and a relatively low BITs score compared the rest which indicates low sequence similarity with *Mus musculus* GULO sequence. The other hits which had lower BITs score and e-value corresponded to dehydrogenase mitochondrial isoforms having different enzymatic properties[26]. From a structural comparison of figure panel 2A and the figure panel 3A, the structure of protein was undoubtedly different from Mus musculus GULO enzyme. This structure, when queried in SMART, did not possess an ALO domain. Furthermore, it did not possess both the HWAK and VGGGHS motif which is required for enzyme catalysis. FAD cofactor binding site was not available for the GULO activity but the FAD domain is present in this reductase enzyme for binding cholesterol[27].

Tblastn, with default parameters, yielded the same best result as blastn which is the GULO pseudogene. When investigating further literature on this result, this sequence was produced artificially through biochemical experiments, which limits our ability to draw concrete conclusions. Other biomedical research supports that GULO gene is nonfunctional in humans; the GULO sequence has shown no particular alternatives with similar sequences which indicates loss of enzymatic activity[28]. Particular exons, when compared to rat GULO gene, that have been retained in this human pseudogene are exons 4, 7, 9, 10 and 12[22][27] which indicates the other critical exons which make up the FAD domain in a functional GULO enzyme are absent. Therefore, with these factors mentioned above, the hu-

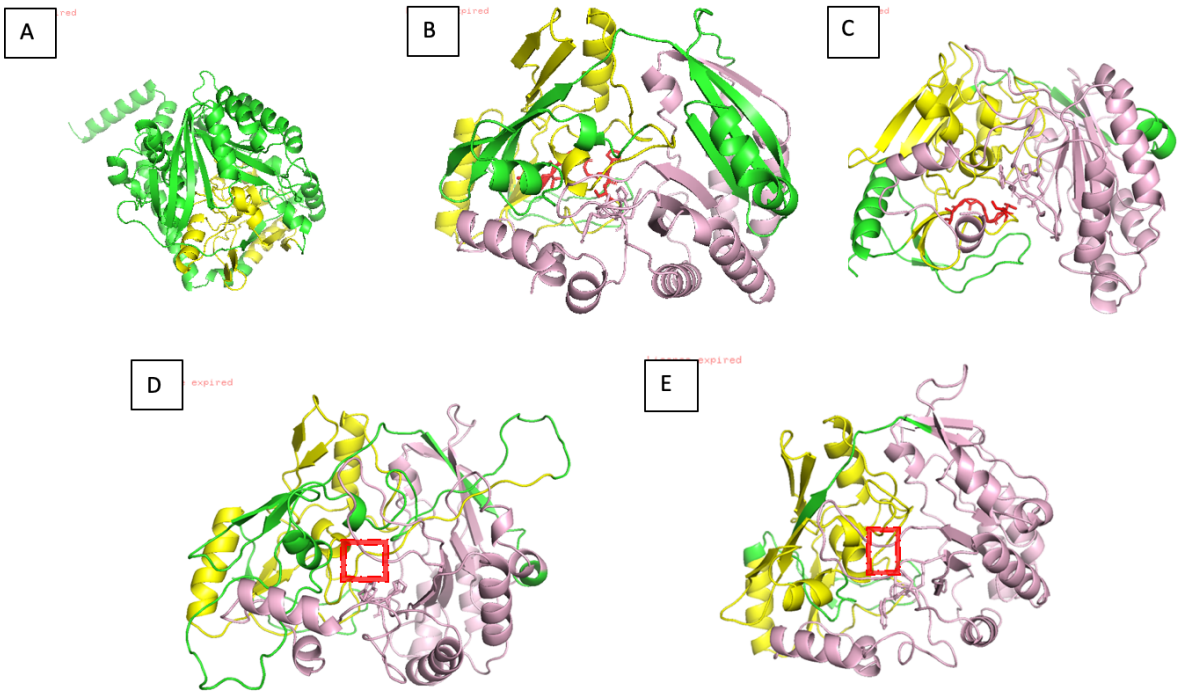

Figure 3: PYMOL-generated figures: FAD domains (yellow) and ALO domains (light pink) A) AlphaFold structure of Delta(24)-sterol reductase precursor (*Homo sapiens*) **B)** AlphaFold predicted structure of L-gulonolactone oxidase, HWAK motif (pink sticks), VGGGHS motif (red sticks) (*Xenopus tropicalis*) C) Modeller-predicted structure of L-gulonolactone oxidase, HWAK motif (pink sticks), VGGGHS motif (red sticks) *Elephas maximus*, **D)** Modeller-predicted structure of *Cavia porcellus*, HWAK motif (pink sticks), missing VGGGHS motif (red box) **E)** Modeller-predicted structure of *Branchiostoma lanceolatum*, HWAK motif (pink sticks), missing VGGGHS motif (red box)

BLAST algo.	Description	Tot. score	BITS score	E-value	Accession no.
blastn	*Homo sapiens* gulonolactone (L)-oxidase pseudogene (GULOP) on chromosome 8	180	180	5e-42	NG_001136.2
blastp	Delta(24)-sterol reductase precursor (*Homo sapiens*)	71	71	1e-11	NP_055577.1
tblastn	*Homo sapiens* gulonolactone (L)-oxidase pseudogene (GULOP) on chromosome 8	68.6	68.6	1e-09	NG_001136.2

Table 1: Best hits based on BITs score and E-value from BLAST algorithms (Blastn, Blastp tBlastn) for *Homo sapiens*

man GULO gene is nonfunctional and exists as a pseudogene.

3.2 *Xenopus tropicalis*

All 3 algorithms (blastn, blastp and tblastn) yielded significant results where all the hits obtained 3 hits which had an e-value of 0. Blastn result was obtained with the loosest parameters with lowest match/mismatch scores and decreased word size which yielded the predicted sequence of L-gulonolactone oxidase (LOC101733882)(accession id: XM_031903103.1). It is not surprising that the BITs score is high at 920 with the parameters loosened. However, it is important to note that this nucleotide sequence was computationally predicted by Gnomon; due to this, it is difficult to derive solid conclusions on functionality.

Blastp hits were obtained with default parameters; most reliable blastp result was L-gulonolactone oxidase(accession id: XP_031758963.1) which is a protein derived from the result yielded in blastn and tblastn. This result showed complete query coverage at 100% and 0 gap penalties. When strengthening the parameters, BLAST also yielded the similar results with relatively small changes in BITs score and e-values stayed constant at 0. Therefore, it is not surprising that when investigating a struc-

tural comparison of figure 2A and figure 3B, the structure of protein L-gulonolactone oxidase (*Xenopus tropicalis*) from blastp showed high structural similarity to *Mus musculus* GULO enzyme. Both FAD and ALO domains are present when *Xenopus tropicalis*' protein was queried on SMART database which was also crosschecked with data provided on UNIPROT database. The domains can be visualized in figure panel 3B. Furthermore, both catalytic motifs: VGGGHS and HWAK motif are present in this protein, in which FAD binding can occur. However, when investigating from literature and NCBI on how this protein was derived, the nucleotide sequence obtained which was translated into *Xenopus tropicalis*' protein was obtained through whole genome shotgun sequencing[29]. Whole genome shotgun sequencing is known to be prone to amino acid sequencing errors[30], which might affect the true amino acid composition of GULO enzyme in *Xenopus tropicalis* yielding inaccurate BLAST results. Therefore, the sequence predicted for this protein may not be authentic.

Bioinformatics research has shown that *X. tropicalis* contains a FAD binding domain and ALO binding domain. However, this was found on a different locus (LOC1001080753) on the chromosome than nucleotide sequence yielded in tblastn and blastn[31]. It was shown through phylogenetic analysis and multiple sequence alignment that *X. tropicalis* and its related species, *Xenopus laevis* revealed a similar divergence to *Mus musculus* GULO gene[31]. It is quite unusual as there has been literature evidence supporting GULO functionality in *X. laevis*[1][31] but not in *X. tropicalis*. There has also been evidence for high chromosomal synteny between GULO sequence in *Mus musculus* and gene record LOC100180753 in *X. tropicalis* genome[31]. However, in this study, X. tropicalis gene was revised as the original sequence was suspected to contain large fragment of repetitive motifs. This study contained artificial manipulation of the species' genome[31]. Therefore, taking all these factors into consideration, it is quite likely that GULO gene in *Xenopus tropicalis* is nonfunctional as all of the best hits from BLAST were computationally predicted.

3.3 *Elephas maximus*

3 BLAST algorithms yielded significant results for *Elephas maximus*. Blastn yielded the best result with PREDICTED: *Elephas maximus indicus* L-gulonolactone oxidase (LOC126065470), mRNA (accession id: XM_049865493.1) which has the highest BIT score and lowest e-value with strong parameters implemented (BLOSUM62 matrix, word size=32). Moreover, tblastn result yielded the same result as blastn with the same high BITs score and e-value of 0. Algorithm parameters implemented were relatively strong (BLOSUM90 matrix, and high gap costs). Query coverage of tblastn result was 99% of the *Mus musculus* nucleotide sequence, where there were minor nucleotide substitutions. With both blastn and tblastn results, it can be hypothesized that blastp would yield a L-gulonolactone oxidase enzyme in this species. Hence, investigation into blastp would give a clearer interpretation into the functionality of GULO gene in this species.

Blastp yielded significant results with L-gulonolactone oxidase of *Elephas maximus* (accession id: XP_049721450.1) which led to a high BITs score, e-value of 0, and 100% query cover. Blastp parameters were the strongest (BLOSUM90 matrix, high gap costs). As 0.90 similarity score criterion was adopted, this indicates GULO gene of *Elephas maximus* was recently divergent from *Mus musculus* GULO gene. Furthermore, pairwise comparison between both of these sequences show high sequence alignment and multiple non-deleterious substitutions as amino acid mutations were of similar biochemical properties. It could be now suspected that the GULO enzyme present in *Elephas maximus* is structurally similar to *Mus musculus* GULO enzyme. Based on structural analysis (Figure panel 3C) and query in SMART database of L-gulonolactone oxidase of *Elephas maximus* indicates that there is the FAD and ALO domain present in this species. The VGGGHS and HWAK motifs are also present for FAD and ALO domains respectively for the catalytic reaction to occur. The enzyme structure of *Elephas maximus* is relatively similar to that of the *Mus musculus* GULO enzyme. This is quite indicative of relatively few amino acid substitutions which are not deleterious to the protein function. Pairwise comparison between *Mus musculus* GULO and *Elephas maximus* GULO on blastp shows evident amino acid substitutions at non-conserved regions of the protein. The structure (Figure 3C) shows that the catalytic region with both motifs are at the interface of both domains which is structurally similar to *Mus musculus* GULO. It can be inferred that through structural comparison with *Mus musculus* GULO enzyme (Figure panel 2A), that *Elephas maximus* has a functional GULO protein. Based on the diet of *Elephas maximus*, their intake of vitamin C is generally low[32]; it has also been reported that *Elephas maximus* has 4 times more vitamin C in their milk than cows[33]. Through phylogenetic analysis, it is also indicated that *Elephas maximus* can synthesize vitamin C[34] as their vitamin C consumption is comparatively lower than species which can synthesize them[34]. Therefore, it is quite indicative that *Elephas maximus* has a functional GULO gene.

BLAST algo.	Description	Tot. score	BITS score	E-value	Accession no.
blastn	PREDICTED: *Xenopus tropicalis* L-gulonolactone oxidase (LOC101733882)	920	920	0.0	XM_031903103.1
blastp	L-gulonolactone oxidase (*Xenopus tropicalis*)	683	683	0.0	XP_031758963.1
tblastn	PREDICTED: *Xenopus tropicalis* L-gulonolactone oxidase (LOC101733882), mRNA	658	658	0.0	XM_031903103.1

Table 2: Best hits based on BITs score and E-value from BLAST algorithms (Blastn, Blastp tBlastn) for *Xenopus tropicalis*

3.4 *Cavia porcellus*

3 BLAST algorithms yielded significant results for *Cavia porcellus*. Blastn yielded significant gene orthologs with the top result being PREDICTED: *Cavia porcellus* L-gulonolactone oxidase-like (LOC100735706), mRNA (accession id: XM_013143314.2) with a relatively high BITs score (743) and e-value of 0. This was achieved with default parameters of blastn; increasing word size did not yield any results. The nucleotide sequence, when compared to query sequence, has shown mismatches which includes nucleotide gaps which could induce a frameshift and affect protein. This is indicated in Nishikimi et. al's sequence comparison with the cDNA of the rat L-gulono-γ-lactone oxidase enzyme where they found mutations in exon regions of the *Cavia porcellus* sequence which more than half led to nonconservative amino acid changes and 3 stop codons[35][28]. Blastp result which was LOW QUALITY PROTEIN: L-gulonolactone oxidase-like (*Cavia porcellus*)(accession id: XP_012998768.1) was obtained through the strongest parameters (BLOSUM90, composition-based statistics adjustment and highest gap costs). However, in the literature cited under blastp result, it was known that this genome was artificially constructed to correct this low-quality protein; 2 bases were inserted in 1 codon and inserted 2 bases at 2 genomic stop codons. Artificial manipulation of nucleotide sequence decreases the reliability of the blastp result as the genome was manipulated and potentially the structure will not indicate the functionality of protein in the natural environment.

When blastp result was queried on SMART database, unsurprisingly with the manipulated sequence, FAD and ALO domains are present. With a structural comparison with *Mus musculus* GULO enzyme (Figure panel 2A) and *Cavia porcellus* L-gulonolactone oxidase-like protein (Figure panel 3D), it is presented that the two critical domains, and the catalytic region are present at the interface of both domains. Aside from being structurally similar to HWAK motif is also present in the ALO domain of this species. However, it does not have the VGGGHS motif required for FAD bind-

BLAST algo.	Description	Tot. score	BITS score	E-value	Accession no.
blastn	PREDICTED: *Elephas maximus indicus* L-gulonolactone oxidase (LOC126065470), mRNA	1459	1459	0.0	XM_049865493.1
blastp	L-gulonolactone oxidase (*Elephas maximus indicus*)	863	863	0.0	XP_049721450.1
tblastn	PREDICTED: *Elephas maximus indicus* L-gulonolactone oxidase (LOC126065470), mRNA	1459	1459	0.0	XM_049865493.1

Table 3: Best hits based on BITs score and E-value from BLAST algorithms (Blastn, Blastp tBlastn) for *Elephas maximus*

BLAST algo.	Description	Tot. score	BITS score	E-value	Accession no.
blastn	PREDICTED: *Cavia porcellus* L-gulonolactone oxidase-like (LOC100735706), mRNA	743	743	0.0	XM_013143314.2
blastp	LOW QUALITY PROTEIN: L-gulonolactone oxidase-like (*Cavia porcellus*)	615	615	0.0	XP_012998768.1
tblastn	*Cavia porcellus* L-gulono-gamma-lactone oxidase gene homologue, exon 2, 3, 4, 5, 7, 8, 9 corresponding parts	95.9	95.9	3e-20	D12762.2

Table 4: Best hits based on BITs score and E-value from BLAST algorithms (Blastn, Blastp tBlastn) for *Cavia porcellus*

ing. Histidine residue in the VGGGHS motif was not identified which is a critical residue required for binding as inferred from how highly conserved the residue is from the *Mus musculus* GULO enzyme shown (Figure panel 2E). This shows more evidence that GULO gene is nonfunctional in *Cavia porcellus*. Tblastn's best hit was *Cavia porcellus* L-gulono-γ-lactone oxidase gene homologue, exon 2, 3, 4, 5, 7, 8, 9 corresponding parts (accession id: D12762.2). These exons corresponded to ones found when compared to rat L-gulonolactone oxidase gene homologue[35]. This is true when analyzing the blastn result which showed major deletions in the guinea pig sequence, which caused the lost of exon 1 and 6 which was essential for both domains[36][37]. Therefore, it can be concluded that the GULO gene is nonfunctional in this species.

3.5 *Branchiostoma lanceolatum*

Blastn did not yield significant results even with the parameters loosened to the lowest (word size: 16). Tblastn, however, yielded significant results with parameters strengthened to the BLOSUM90 and gap costs strengthened. The best hit obtained was *Branchiostoma lanceolatum* genome assembly, chromosome 10 (accession id: OV696695.1) with a relatively low e-value but low BITs score as compared to the other BLAST searches yielded in other species. Query cover of this genome assembly was around 47% which is also relatively low. The other 2 hits obtained had lower e-values and BITs score but were also part of the genome assembly on different chromosomes (chromosome 2 and 8). It is hard to evaluate the functionality of the gene when it is known there is lack of complete genome sequencing and annotation in lancelets[3].

However, blastp obtained significant hits when strengthening the parameters to BLOSUM90. From blastp searches, the protein obtained for *Branchiostoma lanceolatum* was D2HGDH protein (accession id: CAH1255212.1), which has a high BITs score and significantly low e-value close to 0. Query coverage was also relatively high compared to the other hits at 98%. From this result, it is quite indicative that the GULO gene is present in the species. This supports Yang et al's study which indicates that lancelets have 11 exons responsible for the GULO gene within their genome[3]. The other 5 hits were hypothetical proteins derived from *Branchiostoma lanceolatum's* genome assembly, which had lower BITs score and e-values. With a pairwise comparison, there were a few non-deleterious amino acid substitutions where residues

BLAST algo.	Description	Tot. score	BITS score	E-value	Accession no.
blastn	No significant results yielded				
blastp	D2HGDH (*Branchiostoma lanceolatum*)	484	484	1e-169	CAH1255212.1
tblastn	*Branchiostoma lanceolatum* genome assembly, chromosome 10	138	80.7	2e-14	OV696695.1

Table 5: Best hits based on BITs score and E-value from BLAST algorithms (Blastn, Blastp tBlastn) for *Branchiostoma lanceolatum*

mutated were of similar properties; there were also very few mismatches seen in the sequence when compared to Mus musculus amino acid sequence. When D2HGDH protein was queried on SMART database, this protein has both the FAD and ALO domains, which can be shown in figure panel 3E. In terms of a structural comparison with Mus musculus GULO gene, D2HGDH protein is also quite structurally similar to GULO enzyme, due to existence of both domains[36][37]. It also contains the HWAK motif which is required for FAD binding but does not contain the VGGGHS motif and conserved histidine residue which indicates that the GULO enzyme is nonfunctional. From the findings above, functionality of GULO gene in this species remains inconclusive due to lack of genomic data.

4 Discussion

Parameters which yielded many significant results when using default parameters such as *Homo sapiens* BLOSUM90 tblastn, were switched to composition-based statistics for their compositional adjustments. This method was implemented as this can reduce false positive searches in circumstances of hits matching query sequence of different lengths such as shown in genome assemblies curated in species mentioned above where whole genome was presented on NCBI[38]. This refinement yielded more reliable and accurate hits to which functionality was investigated in individual species. It is conclusive that Homo sapiens GULO gene is nonfunctional and exists as a pseudogene. Based on the findings from BLAST results, the multiple mutations which have rendered it nonfunctional in terms of vitamin C synthesis were heavily supported by literature and structural analysis of the blastp result. This is also the sa r *Cavia porcellus* where based on the BLAST lts being all computationally predicted, and literature inferring nonfunctionality of the GULO gene. This was further confirmed through structural analysis where it did not contain VGGGHS motif where the highly conserved residue was not found within this structure. Furthermore, evidence for a nonfunctional GULO gene for *Xenopus tropicalis* with structural analysis and literature research. This was also true with *Elephas maximus* where there are conclusive results on a functional GULO gene based on structural analysis, BLAST results, and literature on their diets. The only species which was rendered inconclusive in terms of findings was *Branchiostoma lanceolatum* where structural analysis has shown VGGGHS motif was absent in the structure and tblastn result yielded a genome assembly which was difficult to analyze.

Further investigation into certain dietary requirements and species environment would provide a better reasoning on how these mutations have occurred over evolutionary time and why certain species have lost GULO functionality. The environment could contribute to the epigenetics of how, over evolutionary time, certain species such as *Homo sapiens* and *Cavia porcellus* has lost its ability to biosynthesize vitamin C. Better genomic data should be generated as well for some of the species discussed above to draw more conclusive results. The method could also be improved if PSI-blast was used or a combinatorial method[39] of global and local alignment algorithms to generate more accurate alignments[39][40]. Smith-Waterman alignment gains more precise results when searching sequence homology than BLAST but is computer intensive and time consuming. PSI-blast is also better than blastp in terms of producing more hits with significant e-values and is useful for finding distantly related proteins[41].

5 References

1. Henriques, S. F. *et al.* Multiple independent L-gulonolactone oxidase (GULO) gene losses and vitamin C synthesis reacquisition events in non-Deuterostomian animal species. *BMC Evolutionary Biology* **19** (June 2019).

2. Maeda, N. *et al.* Aortic wall damage in mice unable to synthesize ascorbic acid. *Proceedings of the National Academy of Sciences* **97,** 841–846. (2022) (Jan. 2000).

3. Yang, H. Conserved or Lost: Molecular Evolution of the Key Gene GULO in Vertebrate Vitamin C Biosynthesis. *Biochemical Genetics* **51,** 413–425 (Feb. 2013).

4. Leferink, N. G., Heuts, D. P., Fraaije, M. W. & van Berkel, W. J. The growing VAO flavoprotein family. *Archives of Biochemistry and Biophysics* **474,** 292–301. `https://pure.rug.nl/ws/files/6719385/2008ArchBiochemBiophysLeferink.pdf` (2022) (June 2008).

5. Aboobucker, S. I. & Lorence, A. Recent progress on the characterization of aldonolactone oxidoreductases. *Plant Physiology and Biochemistry* **98,** 171–185. `https://www.sciencedirect.com/science/article/pii/S0981942815301728` (2022) (Jan. 2016).

6. Forneris, F. *et al.* Structural analysis of the catalytic mechanism and stereoselectivity in Streptomyces coelicolor alditol oxidase. *Biochemistry* **47,** 978–985. `https://pubmed.ncbi.nlm.nih.gov/18154360/` (2022) (Jan. 2008).

7. Rentzsch, R. & Orengo, C. A. Protein function prediction using domain families. *BMC Bioinformatics* **14** (Feb. 2013).

8. Pevsner, J. *Bioinformatics and functional genomics* 107 (Wiley Blackwell, 2015).

9. Ewens, W. J. & Grant, G. R. *Statistical Methods in Bioinformatics* 245 (Springer Science Business Media, Mar. 2013).

10. Pevsner, J. *Bioinformatics and functional genomics* 108–109 (Wiley Blackwell, 2015).

11. Pevsner, J. *Bioinformatics and functional genomics* 110–112 (Wiley Blackwell, 2015).

12. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research* **28,** 231–234. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102444/ (Jan. 2000).

13. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics* **15,** 5.6.1–5.6.30 (Sept. 2006).

14. Alexandrov, V. & Gerstein, M. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics* **5,** 2. (2020) (2004).

15. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33,** W244–W248. (2019) (July 2005).

16. Söding, J. *Protein structure and function prediction by pairwise comparison of hidden Markov models.* https://www.researchgate.net/profile/Johannes-Soeding/publication/239581534_Protein_structure_and_function_prediction_by_pairwise_comparison_of_hidden_Markov_models/links/00b7d53a9080fe1735000000/Protein-structure-and-function-prediction-by-pairwise-comparison-of-hidden-Markov-models.pdf.

17. Kirshner, D. A., Nilmeier, J. P. & Lightstone, F. C. Catalytic site identification—a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Research* **41,** W256–W265. (2022) (May 2013).

18. Duque, P., Vieira, C. P. & Vieira, J. Advances in Novel Animal Vitamin C Biosynthesis Pathways and the Role of Prokaryote-Based Inferences to Understand Their Origin. *Genes* **13,** 1917. (2022) (Oct. 2022).

19. Armon, A., Graur, D. & Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology* **307,** 447–463 (Mar. 2001).

20. Sikosek, T. & Chan, H. S. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society, Interface* **11,** 20140419. https://www.ncbi.nlm.nih.gov/pubmed/25165599 (2014).

21. Kenney, W. C. *et al.* Identification of the covalently-bound flavin of l-galactonolactone oxidase from yeast. *FEBS Letters* **97,** 40–42. https://www.sciencedirect.com/science/article/pii/0014579379800472 (2022) (Jan. 1979).

22. Nishikimisp, M., Fukuyaman, R., Minoshiman, S., Shimizux, N. & Yagis, K. THE JOURNAL OF BIOUX ICAL CHEMISTRY Cloning and Chromosomal Mapping of the Human Nonfunctional Gene for L-Gulono-y-lactone Oxidase, the Enzyme for L-Ascorbic Acid Biosynthesis Missing in Man*. **269,** 13685–13688. https://www.jbc.org/article/S0021-9258(17)36884-9/pdf (1994).

23. Nishikimi, M. & Yagi, K. Molecular basis for the deficiency in humans of gulonolactone oxidase, a key enzyme for ascorbic acid biosynthesis. *undefined.* https://www.semanticscholar.org/paper/Molecular-basis-for-the-deficiency-in-humans-of-a-Nishikimi-Yagi/b58dbf35ab77c06b134ef70ea45fc2864305ae6c (2022) (1991).

24. Nishikimi, M., Koshizaka, T., Ozawa, T. & Yagi, K. Occurrence in humans and guinea pigs of the gene related to their missing enzyme l-gulono--lactone oxidase. *Archives of Biochemistry and Biophysics* **267,** 842–846. (2019) (Dec. 1988).

25. 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. *Journal of Nutritional Science and Vitaminology* **49,** 315–319 (2003).

26. Feng, D. *et al.* Mitochondrial Aldehyde Dehydrogenase 2 Represents a Potential Biomarker of Biochemical Recurrence in Prostate Cancer Patients. *Molecules* **27,** 6000. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9500792/ (2022) (Sept. 2022).

27. Greeve, I. *et al.* The human DIMINUTO/DWARF1 homolog seladin-1 confers resistance to Alzheimer's disease-associated neurodegeneration and oxidative stress. *The Journal of neuroscience* **20,** 7345–7352. https://europepmc.org/article/MED/11007892 (2022) (Oct. 2000).

28. Zhang, Z. D., 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 Biology* **11,** R26. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864566/ (2010).

29. Hellsten, U. *et al.* The genome of the Western clawed frog Xenopus tropicalis. *Science (New York, N.Y.)* **328,** 633–636. https://www.ncbi.nlm.nih.gov/pubmed/20431018 (Apr. 2010).

30. Ferguson, E. *The strengths and weaknesses of whole-genome sequencing* 2020. https://inspirestudentjournal.co.uk/wp-content/uploads/2020/10/Inspire-Student-Journal-Emily-Ferguson.pdf.

31. Xie, Y. *et al.* Gulo Acts as a ¡b¿i¿de novo¡/i¿¡/b¿ Marker for Pronephric Tubules in ¡b¿i¿Xenopus laevis¡/i¿¡/b¿. *Kidney and Blood Pressure Research* **41,** 794–801 (2016).

32. Ruetten, M. *et al.* Iron Regulation in Elderly Asian Elephants (Elephas maximus) Chronically Infected With Mycobacterium tuberculosis. *Frontiers in Veterinary Science* **7.** (2022) (Oct. 2020).

33. Dhairykar, M. & Singh, K. Management of nutrition in captive Asian elephants. *160 International Journal of Veterinary Sciences and Animal Husbandry* **5,** 160–163. https://www.veterinarypaper.com/pdf/2020/vol5issue4/PartC/5-4-31-256.pdf (2020).

34. Drouin, G., Godin, J.-R. & Page, B. The Genetics of Vitamin C Loss in Vertebrates. *Current Genomics* **12,** 371–378. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3145266/ (Aug. 2011).

35. Nishikimi, M., Kawai, T. & Yagi, K. Guinea pigs possess a highly mutated gene for L-gulono-gamma-lactone oxidase, the key enzyme for L-ascorbic acid biosynthesis missing in this species. *Journal of Biological Chemistry* **267,** 21967–21972. (2021) (Oct. 1992).

36. Cui, J., Pan, Y.-H., Zhang, Y., Jones, G. & Zhang, S. Progressive Pseudogenization: Vitamin C Synthesis and Its Loss in Bats. *Molecular Biology and Evolution* **28,** 1025–1031. (2020) (Oct. 2010).

37. BURNS, J. J., PEYSER, P. & MOLTZ, A. Missing Step in Guinea Pigs Required for the Biosynthesis of L-Ascorbic Acid. *Science* **124,** 1148–1149. (2019) (Dec. 1956).

38. Pevsner, J. *Bioinformatics and functional genomics* 111 (Wiley Blackwell, 2015).

39. Polyanovsky, V. O., Roytberg, M. A. & Tumanyan, V. G. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms for Molecular Biology* **6,** 25 (2011).

40. Wieds, G. *Bioinformatics Bioinformatics Explained Explained Bioinformatics explained: BLAST versus Smith-Waterman* 2007. https://people.montefiore.uliege.be/kvansteen/GBIO0009-1/ac20092010/Class4/BE-smith-waterman_versus_blast.pdf.

41. Pevsner, J. *Bioinformatics and functional genomics* 142 (Wiley Blackwell, 2015).