Introduction

L-gulono-gamma-lactone oxidase (GULO) is an enzyme used to synthesize vitamin C. Vitamin C is biosynthesized from glucose in the liver or kidneys of animals, with the final step catalyzed by L-gulono-gamma-lactone oxidase. [1] GULO is an old gene that predates the separation of Animals and Fungi, although it could be much older. GULO is a microsomal enzyme that catalyzes the aerobic conversion of gulono-lactone to ascorbic acid and the production of hydrogen peroxide. [2] Invertebrates, bony fish, some finches, guinea pigs, bats and hominid primates, including humans, are deficient in functional GULO and therefore require dietary intake of ascorbic acid as vitamin C. Animal species where the GULO gene has been lost usually retain the remaining genes of the pathway.[3]

To explore the evolution and absence of this gene between these species and whether the functionality of the protein encoded by the gene is affected, this experiment provides a Mouse DNA coding sequence for the L-gulonolactone oxidase gene and match it with five species of vertebrates which are Homo sapiens, Xenopus tropicalis, Elephas maximus, Cavia porcellus and Branchiostoma lanceolatum. By sequence matching using BLASTN, BLASTP and TBLASTN algorithms, this experiment will analyse its proteins and genes and decide whether these species have a functional GULO gene. Furthermore, changing different parameters and different databases will help this experiment achieve its objectives better.

Materials and methods

The experiment requires the Mouse DNA coding sequence for the L-gulonolactone oxidase gene and sequence of the protein encoded by this gene and query length is1323. By searching NCBI BLASTX, gene sequence could match corresponding protein sequence which NCBI Reference Sequence is NP_848862.1. BLASTX translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each of these proteins to a protein database.[4]

Sequence matching analysis using the Blast algorithm. BLAST is a heuristic algorithm for finding comparable sequences in large databases and is an approximate comparison algorithm based on local comparisons, which can greatly reduce the running time of a program while maintaining high accuracy. It is a local matching-based approximation algorithm that can significantly reduce the running time of a program while maintaining high accuracy.

The basic principle of the BLAST algorithm is to improve the speed of comparison by generating a smaller number but better quality of augmented points. The principle of the algorithm is mainly divided into the following five steps: (1) Filtering: first filter out the low complexity region, (2) Seeding: combine every k words in the Query sequence into a table, (3) Comparing: list all possible word groups, then with the replacement matrix to give the high score word groups and organize them into a fast search tree structure or index, so this step can quickly search all the words in the extensive data set Matching sequences, find the position of each seed word in the reference sequence; (4) extension: when the position of the seed word is found, the subsequent need to extend the seed word into a long fragment, extension process, the score value is also changing, when the score value is less than a threshold value that stops extension, the final fragment to become a high score fragment pairs, (5) significance analysis, and finally use the the formula calculates the E value, which

measures the number of pairs that exist in the database with better match scores than the current one in a random situation, so the value can be used as an indicator to evaluate the credibility of the HSP pair sequences.

This experiment applies the following three main algorithms: BLASTN, BLSATP and TBLASTN. BLASTN is a standard search, a query of nucleic acid sequences to a nucleic acid library, where each known sequence present in the library is compared one-to-one with the query sequence. BLASTP is also a standard search, a query for protein sequences to a protein sequence database. TBLASTN was used to identify database sequences encoding proteins which is similiar to the query. The nucleic acid database sequences were first translated, after which the compiled protein was compared with the protein sequence to be queried.

In the choice of database, the Nucleotide collection (nr/nt) Database and Non-redundant protein sequences(nr) were used in this experiment. The Non-Redundant Protein Sequence Database is a non-redundant protein library containing all non-redundant protein sequences from GenBank+EMBL+DDBJ+PDB; for all known or probable coding sequences, the corresponding amino acid sequence is given in the NR record as well as the sequence number in the specialized protein database.[5] Library corresponds to a cross-index based on nucleic acid sequences, linking nucleic acid data to protein data: the nucleic acid sequence database, a subset of the NR library. The default algorithm parameters for this experiment are as follows:

1. BLASTN: max target sequences are 100, expect threshold is 0.05, word size is 28, max matches in a query range is 0, match/mismatch scores is 2,-3, gap costs is linear, low complexity regions and mask for lookup table only.
2. BLASTP, TBLASN: max target sequences are 100, expect threshold is 0.05, word size is 6, max matches in a query range is 0, the matrix is BLOSUM62, gap costs: existence:11 extension:1 and compositional adjustments: conditional compositional score matrix adjustment.

When the query result does not meet expectations, searching for relatives of the species, searching genome-wide database, and changing parameters like increasing expected threshold will be used to enrich the results and judge the conclusions of the experiments.

Results

Homo sapiens

In the Nucleotide collection (nr/nt) database, BLASTN is used to match the Mouse DNA coding sequence for the L-gulonolactone oxidase gene with the Homo sapiens to search for somewhat similar sequences. The results show 12 chosen sequences, and there is a pseudogene (GULOP) on Homo sapiens chromosome 8, which is the best matched sequence, and its sequence id is NG_001136.2, sequence length is 11663, and the number of matches is 4. Compared with the mouse GULO sequence, the sequence fragment with the highest similarity between the two appear in this sequence from 9437 to 9599, which its score is 189 bits, and there are 141 pairs of bases in the sequence fragment with a total length of 164 that can be matched. Furthermore, there is a gap between them. The whole sequence score is 529 points, the coverage of the target sequence is 40%, the expected value is 2e-44, and the sequence similarity is 85.98%. The query sequence is not an exact match for the subject sequence, and there are base mismatches and base deletions in between. In taxonomy, the

number of hits is 12.
When using the mouse's GULO protein sequence, which NCBI Reference Sequence is NP_848862.1 to match with homo sapiens in the non-redundant protein sequences, the results show no significant similarity found. After trying change parameters like increasing expect threshold and choosing organism as Home (taxid:9605), there is also not any BLASTP result.
It was then using TBLASTN to match the mouse's GULO protein sequence with the homo sapiens nucleic acid database in the Nucleotide collection (nr/nt). There are five matched results, and the sequence with the highest match is also what is found in BLASTN, which sequence id is NG_001136.2 and is a pseudogene (GULOP) on Homo sapiens chromosome 8. Its max score is 68.6, and its total score is 172, its query cover is 25%, its E-value is 1e-09, and its sequence similarity is 77.5%. In this sequence, its sequence length is 11663, the number of matches is 3. In taxonomy, the number of hits is 5.
Also, Homo sapiens GULO gene for L-gulono-gamma-lactone oxidase, which sequence id is D17461.1, its sequence length is 3364, and its number of Matches is two also has the same E-value in TBLASTN result. Its max score is 68.6, and its total score is 121, its query cover is 17%, its E-value is 1e-09, and its sequence similarity is 77.5%. There are protein mismatches and deletions in both sequences.
The results of the three analyses show that homo sapiens lack a functional GULO gene but have a pseudogene (GULOP) on Homo sapiens chromosome 8 and a Homo sapiens GULO gene for L-gulono-gamma-lactone oxidase.

Xenopus tropicalis

Matching mouse DNA coding sequence for the L-gulonolactone oxidase gene with the Xenopus tropicalis (taxid:8364) and Xenopus tropicalis group (taxid:8363) in the Nucleotide collection (nr/nt) database by using BLASTN, the same result shows that there are two matched results and Xenopus tropicalis L-gulonolactone oxidase match with it which is the best match and its sequence id is XM_031903103.1, sequence length is 1533, and the number of matches is 1. Its max score is 737, its total score is 737, its query cover is 98%, its E-value is 0, and its sequence similarity is 72.68%. It has 20 gaps which are 1% of the total. In taxonomy, the number of hits is 2. There are base mismatches and base deletions in between query sequence and subject sequence.
Using the previous protein sequence to match with Xenopus tropicalis (taxid:8364) and Xenopus tropicalis group (taxid:8363) protein sequence in the non-redundant protein sequences(nr), it has the same result and shows that Xenopus tropicalis L-gulonolactone oxidase is the only result. Its max score is 683, and its total score is 638, its query cover is 100%, its E-value is 0, and its sequence similarity is 71.82%. Furthermore, its sequence id is XP_031758963.1, the sequence length is 440, and the number of matches is 1. It does have any gap. In taxonomy, the number of hits is 2.
When matching mouse's GULO protein sequence with Xenopus tropicalis (taxid:8364) and Xenopus tropicalis group (taxid:8363) nucleic acid database in the Nucleotide collection (nr/nt). Also, the database shows the same sequence, which is Xenopus tropicalis L-gulonolactone oxidase. Its max score is 658, and its total score is 658, its query cover is 100%, its E-value is 0, and its sequence similarity is 69,55%. At the same time, its sequence id

is XM_031903103.1, sequence length is 1533, and the number of matches is 1. In taxonomy, the number of hits is 2.
Because Xenopus tropicalis have L-gulonolactone oxidase protein and gene by BLASTN, BLASTP and TBLASTN, the species was identified as having a functional GULO gene.

Elephas maximus

In the Nucleotide collection (nr/nt) database, BLASTN matches the Mouse DNA coding sequence for the L-gulonolactone oxidase gene with the Elephas maximus (taxid:9783) and Elephas (taxid:9782) to search for somewhat similar sequences. It shows that there is no significant similarity found. Then by using BLASTP and TBLASTN, it also shows that there is no significant similarity found. Trying another way is to change parameters that change the expected threshold from 0.05 to 1; it also shows that there is no significant similarity found.
When changing database from Nucleotide collection (nr/nt) database to Whole-genome shotgun contigs(wgs), BLASTN results have seven matches which the best one is Elephas maximus isolate Icky scaffold_602. Its sequence id is JABTCH010000323.1, its sequence length is 1707894, and the number of Matches is 1. Furthermore, the best match is from 388059 to 388227. Its max score is 268, and its total score is 268, its query cover is 12%, its E-value is 5e-69, and its sequence similarity is 95.27%.
Similarly, in the Whole-genome shotgun contigs(wgs), TBLASTN result shows the best match is the Elephas maximus isolate Icky scaffold_602, which is the same as the BLASTN result. Its max score is 121, and its total score is 881, its query cover is 95%, its E-value is 2e-27, and its sequence similarity is 82.19%.

Cavia porcellus

Matching mouse DNA coding sequence for the L-gulonolactone oxidase gene with the Cavia porcellus (taxid:10141) in the Nucleotide collection (nr/nt) database by using BLASTN. Results show that a predicted Cavia porcellus L-gulonolactone oxidase-like is the best match. Compared with mouse GULO sequence, the sequence fragment with the highest similarity between the two appear in this sequence from 755 to 1474, which its score is 821 bits and there are 614 pairs of bases in the sequence fragment with a total length of 720 that can be matched. The sequence similarity is 85%, the total score is 1231, its query cover is 86% and its E-value is 0. In taxonomy, the number of hits is 2. There are base mismatches in the sequence.
When using BLASTP to match mouse GULO protein sequence with Cavia porcellus (taxid:10141) protein sequence in the non-redundant protein sequences(nr). It shows the only result is that L-gulonolactone oxidase-like low quality protein. Its sequence id is XP_012998768.1, its sequence length is 491 and the number of matches is 1. The best match is from 4 to 491 in the whole sequence which its max score is 615, its total score is 615, its query cover is 99%, its E-value is 0 and its sequence similarity is 65.65%. In taxonomy, the number of hits is 1.
Then matching mouse's GULO protein sequence with Cavia porcellus (taxid:10141) nucleic acid database in the Nucleotide collection (nr/nt) by using TBLASTN, it shows that there are two result, and the best match is a predicted Cavia porcellus L-gulonolactone oxidase-like which its max score is 578, its total score is 578, its query cover is 99%, its E-value is 0 and its

sequence similarity is 62.60%. And its sequence id is XM_013143314.2, its sequence length is 1616 and the number of matches is 1. It has 57 gaps which is 11% of the total and in taxonomy, the number of hits is 2.
Cavia porcellus have L-gulonolactone oxidase-like gene and L-gulonolactone oxidase-like low quality protein by BLASTN, BLASTP and TBLASTN, but it is a predicted gene and low-quality protein so Cavia porcellus cannot be determined whether have a functional GULO gene.

Branchiostoma lanceolatum

It is matching mouse DNA coding sequence for the L-gulonolactone oxidase gene with the Cavia porcellus (taxid:10141) in the Nucleotide collection (nr/nt) database by using BLASTN. Results show that a predicted Cavia porcellus L-gulonolactone oxidase-like is the best match. Compared with the mouse GULO sequence, the sequence fragment with the highest similarity between the two appear in this sequence from 755 to 1474, which its score is 821 bits, and there are 614 pairs of bases in the sequence fragment with a total length of 720 that can be matched. The sequence similarity is 85%, the total score is 1231, its query cover is 86%, and its E-value is 0. In taxonomy, the number of hits is 2. There are base mismatches in the sequence.
When using BLASTP to match mouse GULO protein sequence with Cavia porcellus (taxid:10141) protein sequence in the non-redundant protein sequences(nr). It shows the only result is that L-gulonolactone oxidase-like low-quality protein. Its sequence id is XP_012998768.1, its sequence length is 491, and the number of matches is 1. The best match is from 4 to 491 in the whole sequence which a max score is 615, and its total score is 615, its query cover is 99%, its E-value is 0, and its sequence similarity is 65.65%. In taxonomy, the number of hits is 1.
Then matching mouse's GULO protein sequence with Cavia porcellus (taxid:10141) nucleic acid database in the Nucleotide collection (nr/nt) by using TBLASTN, it shows that there are two result and the best match is a predicted Cavia porcellus L-gulonolactone oxidase-like which its max score is 578, its total score is 578, its query cover is 99%, its E-value is 0, and its sequence similarity is 62.60%. Furthermore, its sequence id is XM_013143314.2, its sequence length is 1616, and the number of matches is 1. It has 57 gaps which are 11% of the total, and in taxonomy, the number of hits is 2.
Cavia porcellus have L-gulonolactone oxidase-like gene and L-gulonolactone oxidase-like low-quality protein by BLASTN, BLASTP and TBLASTN, but it is a predicted gene and low-quality protein, so Cavia porcellus cannot be determined whether it has a functional GULO gene.

BLAST result clearly shows as following:

BLASTN	Max Score	Total Score	Query Cover	E value	Per.Ident	Acc.Len	Accession
Homo sapiens	189	529	40%	2e-44	85.98%	111663	NG_001136.2
Xenopus tropicalis	737	737	98%	0.0	72.86%	1533	XM_031903103.1

Cavia porcellus	821	1231	86%	0.0	85.28%	1616	XM_013143314.2
Cavia porcellus	30.1	30.1	1%	0.98	90.48%	796	FM864151.1

BLASTP	Max Score	Total Score	Query Cover	E value	Per.Ident	Acc.Len	Accession
Xenopus tropicalis	683	683	100%	0.0	71.82%	440	XP_031758963.1
Cavia porcellus	615	615	99%	0.0	65.65%	491	XP_012998768.1

Genome BLASTN	Max Score	Total Score	Query Cover	E value	Per.Ident	Acc.Len	Accession
Elephas maximus	269	1678	99%	2e-69	95.27%	1707894	JABTCH010000323.1
Branchiostoma lanceolatum	57.2	112	13%	4e-06	73.47	8797521	FLLO01000003.1

Genome TBLASTN	Max Score	Total Score	Query Cover	E value	Per.Ident	Acc.Len	Accession
Elephas maximus	121	881	95%	2e-27	82.19%	1707894	JABTCH010000323.1
Branchiostoma lanceolatum	81.6	139	47%	5e-15	29.01%	3300416	FLLO01000023.1

TBLASTN	Max Score	Total Score	Query Cover	E value	Per.Ident	Acc.Len	Accession
Homo sapiens	68.6	121	17%	1e-09	77.50%	3364	D17461.1
Homo sapiens	68.6	172	25%	1e-09	77.50%	11663	NG_001136.2
Xenopus tropicalis	658	658	100%	0.0	69.55%	1533	XM_031903103.1
Cavia porcellus	578	578	99%	0.0	62.60%	1616	XM_013143314.2

Discussion

The GULO protein sequence was not found in the Homo sapiens protein database but had a a pseudogene (GULOP) on Homo sapiens chromosome 8 and a Homo sapiens GULO gene for L-gulono-gamma-lactone oxidase in the Nucleotide collection (nr/nt) database. It means that homo sapiens lack a functional GULO gene but have a pseudogene. It is the same as Yoko

INAI said: “L-Gulono-γ-lactone oxidase (GULO), which catalyzes the last step of ascorbic acid biosynthesis, is missing in humans”.[6]

Following the BLAST result, Xenopus tropicali have L-gulonolactone oxidase protein and gene by BLASTN, BLASTP and TBLASTN. Furthermore, there are some studies have focused on the function of Gulo in the development of the Xenopus pronephros.[7] It shows that gulo, the key gene for vitamin C biosynthesis, can be expressed maternally. This implies the ability to produce endogenous vitamin C in the early egg and zygote blastomere stages. [7] These two species belong to the same species but differ only in their habitat. It means the species was identified as having a functional GULO gene.

Cavia porcellus can be found that they have L-gulonolactone oxidase-like gene and L-gulonolactone oxidase-like low-quality protein by BLASTN, BLASTP and TBLASTN, but it is a predicted gene and low-quality protein, so Cavia porcellus cannot be determined whether it has a functional GULO gene. However, “The capacity to biosynthesize ascorbic acid has been lost in several species including primates, guinea pigs, teleost fishes, bats, and birds.[8] This inability results from mutations in the GLO gene coding for L-gulono-γ-lactone oxidase, the enzyme responsible for catalyzing the last step in the vitamin C biosynthetic pathway” as Marc Y. Lachapelle said, it means Cavia porcellus does not have a functional GULO gene.[8]

Elephas maximus and Branchiostoma lanceolatum all do not have any significant similarities found, and no accurate judgement can be made from the results. Branchiostoma lanceolatum is a cartilaginous fish. However, it has now been established that all cartilaginous and non-teleost bony fish species are able to synthesize vitamin C and that no teleost fish species can do so. [9] So Branchiostoma lanceolatum has a functional GULO gene.

As for Elephas maximus, there is no direct research on whether Elephas maximus has a functional GULO gene. Furthermore, Lara Hasan said: “The absence of L-ascorbic acid (L-AA, or AA) synthesis in scurvy-prone organisms, including humans, other primates, guinea pigs, and flying mammals, was traced to the lack of L-gulonolactone oxidase (GULO) activity” [10] Many mammals have lost the ability to synthesize vitamin C, so perhaps elephants are too but it is not certain. It means Elephas maximus whether having a functional GULO gene cannot be decided now.

Reference

[1] Lee S, Cho M K, Jung J W, et al. Exploring protein fold space by secondary structure prediction using data distribution method on Grid platform[J]. Bioinformatics, 2004, 20(18): 3500-3507.

[2] Ching B, Ong J L Y, Chng Y R, et al. L‐gulono‐7‐lactone oxidase expression and vitamin C synthesis in the brain and kidney of the African lungfish, Protopterus annectens[J]. The FASEB Journal, 2014, 28(8): 3506-3517.

[3] Henriques S F, Duque P, López-Fernández H, et al. Multiple independent L-gulonolactone oxidase (GULO) gene losses and vitamin C synthesis reacquisition events in non-Deuterostomian animal species[J]. BMC evolutionary biology, 2019, 19(1): 1-12.

[4] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool[J]. Journal of molecular biology, 1990, 215(3): 403-410.

[5] Lee S, Cho M K, Jung J W, et al. Exploring protein fold space by secondary structure prediction using data distribution method on Grid platform[J]. Bioinformatics, 2004, 20(18): 3500-3507.
[6] Inai Y, Ohta Y, Nishikimi M. The Whole Structure of the Human Nonfunctional L-Gulono-γ-Lactone Oxidase Gene-the Gene Responsible for Scurvy-and the Evolution of Repetitive Sequences Thereon[J]. Journal of nutritional science and vitaminology, 2003, 49(5): 315-319.
[7] Xie Y, Liu Y, Zhao Y, et al. Gulo Acts as a de novo Marker for Pronephric Tubules in Xenopus laevis[J]. Kidney and Blood Pressure Research, 2016, 41(6): 794-801.
[8] Lachapelle M Y, Drouin G. Inactivation dates of the human and guinea pig vitamin C genes[J]. Genetica, 2011, 139(2): 199-207.
[9] Drouin G, Godin J R, Pagé B. The genetics of vitamin C loss in vertebrates[J]. Current genomics, 2011, 12(5): 371-378.
[10] Hasan L, Vögeli P, Stoll P, et al. Intragenic deletion in the gene encoding L-gulonolactone oxidase causes vitamin C deficiency in pigs[J]. Mammalian genome, 2004, 15(4): 323-333.