UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECB3213 - 01

BIOINFORMATICS DATABASE

SEMESTER 2 2022/2023

GROUP PROJECT

LECTURER:
DR NOOR HIDAYAH BINTI ZAKARIA

SUBMITTED BY:
GROUP 6

GROUP MEMBERS:

Name	**Matric No**
Chong Kah Wei	A20EC0027
Heong Yi Qing	A20EC0043
Lee Jia Yee	A20EC0063

Table of Content

1. Introduction

This project is conducted to design a bioinformatics analytical solution that incorporates the retrieval and processing of bioinformatics data. The goals of this project are to design and construct a thorough procedure for accessing own storage repositories for data storage and analysis as well as to comprehend the significance of data analysis in bioinformatics.

A protein sequence is represented by a group of letters that are arranged from the amino-terminal to the carboxyl-terminal of the protein. To gain the information of protein sequence, it is available through a variety of protein databases including RefSeq, SwissProt, PIR, Genbank, PDB, TPA, UniProt and NCBI. An analysis of protein sequence is the process of examining a protein or peptide sequence to learn more about its features, function, structure or evolution.

Pairwise sequence alignment is one of the commonly used analysis methods to study protein sequences. EMBL-EBI states that pairwise sequence alignment is deployed to determine similarity and show the functional, structural and evolutionary links between two biological sequences of organisms. For multiple sequence alignment (MSA), it is employed with at least three or more biological sequences. MSA has the same ability as pairwise alignment to recognise similarities between organisms and reveal their evolutionary relationship and connection. The current existence of several MSA tools including MUSCLE, MAFFT, T-COFFEE, Kalign, ClustalW and Clustal Omega.

According to the project, the studied protein, human insulin is obtained from the National Center for Biotechnology Information (NCBI). By exploring and investigating the insulin's characteristics between human and four other organisms which are Pongo abelii, Chlorocebus sabaeus, Nycticebus coucang and Apodemus sylvaticus, Clustal Omega algorithm is opted to execute the multiple sequence alignment in R programming.

2. Literature Review

2.1 Review of NCBI Database

National Center for Biotechnology Information (NCBI) is selected as the protein sequence database in the project. The structure of protein data and other attributes stored in NCBI are depicted in Table 1. There are 25 attributes applied to store protein data in NCBI.

Attributes	Description
Accession	The accession number assigned by NCBI.
All fields	All terms from all search fields in the database.
Author	All authors from all references in the records.
EC/RN number	Enzyme Commission (EC) number for an enzyme activity.
Feature key	Biological features listed in the Feature Table of the sequence records.
Filter	Filtered subsets of the database.
Gene name	Gene names annotated on database records.
Genome project	Numeric unique identifier for the genome project that produced the sequence records.
Issue	Issue number of the journals cited on sequence records
Journal	Name of the journals cited on sequence records.
Keyword	Keywords applied by submitter or from controlled vocabularies applied by NCBI or other databases.
Modification date	Date of most recent modification of a sequence record.
Molecular weight	Molecular weight in Daltons of the protein

	chain calculated from the amino acids only.
Organism	Scientific and common names for the complete taxonomy of organisms that are the source of the sequence records
Page number	Page numbers of the articles that are cited on the sequence record.
Properties	Molecular type, source database, and other properties of the sequence record.
Protein name	Names of protein products as annotated on sequence records.
Publication date	Date that records were made public in Entrez.
SeqID string	NCBI identifier string for the sequence record.
Sequence length	Total length of the amino acid sequence
Substance name	Names of chemical substances associated with a record.
Text word	Text on a sequence record that is not indexed in other fields.
Title	Words and phrases found in the title of the sequence record.
Volume	Contains the volume number of the journals in references on the sequence record.
FASTA	A text-based format that stores the amino acid sequences in single-letter codes.

Table 1: Structure of protein data and other attributes stored in NCBI

2.2 Application of Protein Sequence in Multiple Sequence Alignment

A previous research has successfully used multiple sequence alignment to detect 64 recurrent mutations in Coronavirus from Indian isolates. The aim of this study was to investigate mutational sites in the genome of SARS-CoV-2 genomes of Indian isolates by aligning several sequences and comparing them to the reference sequence of Wuhan SARS-CoV-2. (Yashvardhini and Kumar Jha, 2020)

From the NCBI virus database, SARS-CoV-2 ORF1ab polyprotein with the length of 7096 amino acid were retrieved. The sequences from India were chosen and used for further investigation. In order to serve as a reference for mutational analyses, Wuhan type ORF1ab polyprotein was also retrieved. (Yashvardhini and Kumar Jha, 2020)

Using the web service CLUSTAL Omega which combines alignment with HMM profiling, the obtained sequences were aligned. To find the variation in the alignment, Wuhan type SARS-CoV-2 was utilised as the reference. The aligned files were browsed through Jalview and the accession number was noted to identify the alteration in the Indian SARS-CoV-2 protein sequence with regard to the Wuhan type SARS-CoV-2. The studied outcome was SARS-CoV-2 Indian samples had a total of 64 recurrent mutations found in them. (Yashvardhini and Kumar Jha, 2020)

In addition, a phylogenetic tree was generated in the application of CLUSTAL Omega employing the aligned data to determine the relationships between the reference isolate and other individual isolates. As compared to Wuhan SARS-CoV-2, these coronavirus variants grouped in a distinct subtree, demonstrating their diversity and rapid change over time. (Yashvardhini and Kumar Jha, 2020)

To conclude, with the aid of bioinformatics application, CLUSTAL Omega, multiple sequence alignment can ascertain the location of mutations in coronavirus. The creation of a successful vaccine and medication against COVID-19 is facilitated by the occurrence of novel mutations detected at various sites which boosts the viral evolvability.

3. Methodology

There are a total of 3 phases in the project. The first phase is the review phase which covers the review of the protein sequence database and the application of protein sequence in multiple sequence alignment. According to our studies, NCBI can be one of the most famous protein databases that are widely used nowadays. The structure of the protein data and the attributes of this database are analysed. Furthermore, the application of protein sequence in multiple sequence alignment is studied as well in order to gain an insight into the methods to analyse the protein sequence data.

The second phase is the implementation of the protein sequence data in multiple sequence alignment. Basically, this phase consists of two main activities, that is, data collection and implementation of ClustalOmega for protein sequence alignment. The selected protein sequence is human insulin (AAA59172.1) which is obtained from the protein database of National Center Bioinformatics Institution (NCBI). The protein sequence data for multiple sequence alignment are identified and collected using BLASTP and the application of multiple sequence alignment is performed through the application of ClustalOmega algorithm in R language using R studio.

The third phase focuses on result analysis and discussion. Basically, the result of multiple sequence alignment which has been obtained from phase 2 is further analysed and explained in this phase. Other than that, the evolutionary relationship between the proteins will also be illustrated using a phylogenetic tree and further discussed in this phase.

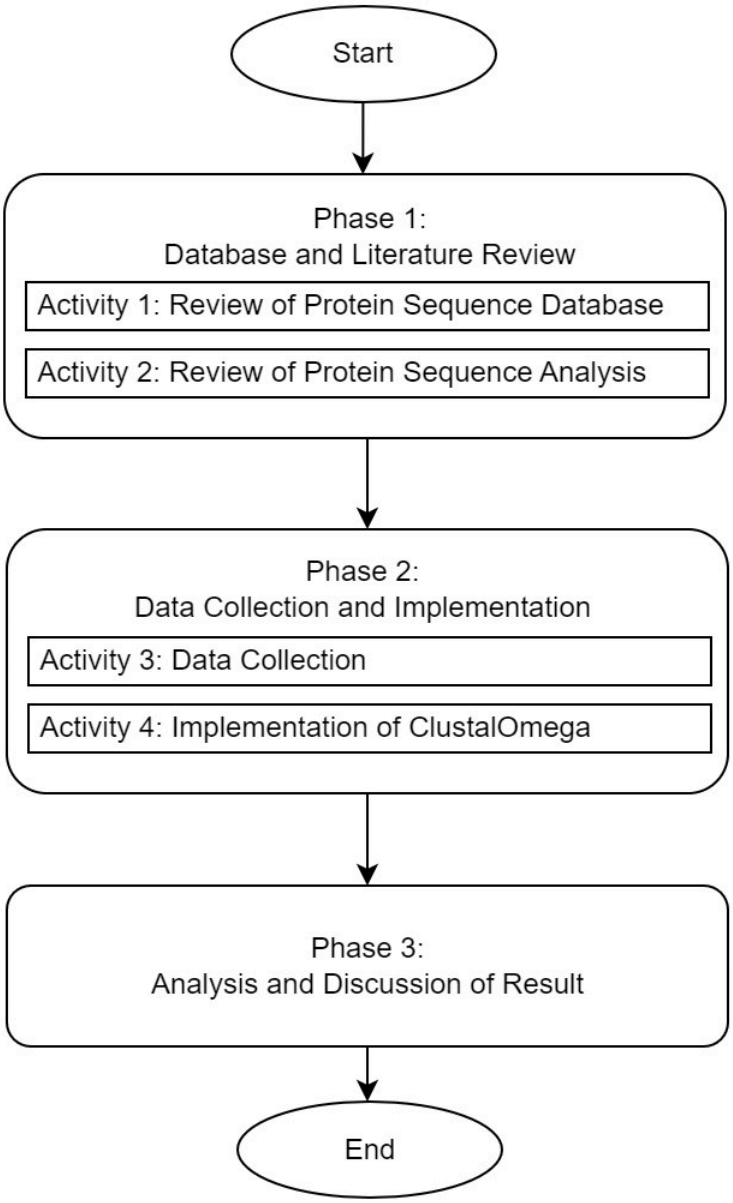
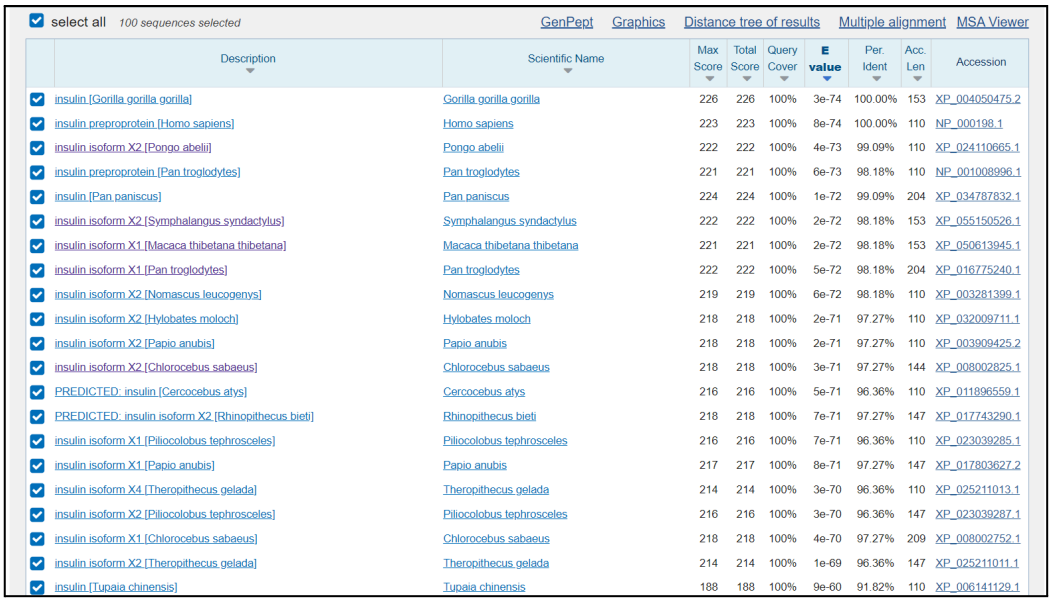


Figure 1: Research Framework Illustration

4. Implementation

4.1 Data Collection

In this project, human insulin is selected in order to carry out the protein sequence analysis. The accession number of human insulin is obtained from NCBI, that is, AAA59172.1. The BLASTP search is then performed using the selected human insulin as a query protein. The search database is restricted to refseq_protein.



select all 100 sequences selected | GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
insulin [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	226	226	100%	3e-74	100.00%	153	XP_004050475.2
insulin preproprotein [Homo sapiens]	Homo sapiens	223	223	100%	8e-74	100.00%	110	NP_000198.1
insulin isoform X2 [Pongo abelii]	Pongo abelii	222	222	100%	4e-73	99.09%	110	XP_024110665.1
insulin preproprotein [Pan troglodytes]	Pan troglodytes	221	221	100%	6e-73	98.18%	110	NP_001008996.1
insulin [Pan paniscus]	Pan paniscus	224	224	100%	1e-72	99.09%	204	XP_034787832.1
insulin isoform X2 [Symphalangus syndactylus]	Symphalangus syndactylus	222	222	100%	2e-72	98.18%	153	XP_055150526.1
insulin isoform X1 [Macaca thibetana thibetana]	Macaca thibetana thibetana	221	221	100%	2e-72	98.18%	153	XP_050613945.1
insulin isoform X1 [Pan troglodytes]	Pan troglodytes	222	222	100%	5e-72	98.18%	204	XP_016775240.1
insulin isoform X2 [Nomascus leucogenys]	Nomascus leucogenys	219	219	100%	6e-72	98.18%	110	XP_003281399.1
insulin isoform X2 [Hylobates moloch]	Hylobates moloch	218	218	100%	2e-71	97.27%	110	XP_032009711.1
insulin isoform X2 [Papio anubis]	Papio anubis	218	218	100%	2e-71	97.27%	110	XP_003909425.2
insulin isoform X2 [Chlorocebus sabaeus]	Chlorocebus sabaeus	218	218	100%	3e-71	97.27%	144	XP_008002825.1
PREDICTED: insulin [Cercocebus atys]	Cercocebus atys	216	216	100%	5e-71	96.36%	110	XP_011896559.1
PREDICTED: insulin isoform X2 [Rhinopithecus bieti]	Rhinopithecus bieti	218	218	100%	7e-71	97.27%	147	XP_017743290.1
insulin isoform X1 [Piliocolobus tephrosceles]	Piliocolobus tephrosceles	216	216	100%	7e-71	96.36%	110	XP_023039285.1
insulin isoform X1 [Papio anubis]	Papio anubis	217	217	100%	8e-71	97.27%	147	XP_017803627.2
insulin isoform X4 [Theropithecus gelada]	Theropithecus gelada	214	214	100%	3e-70	96.36%	110	XP_025211013.1
insulin isoform X2 [Piliocolobus tephrosceles]	Piliocolobus tephrosceles	216	216	100%	3e-70	96.36%	147	XP_023039287.1
insulin isoform X1 [Chlorocebus sabaeus]	Chlorocebus sabaeus	218	218	100%	4e-70	97.27%	209	XP_008002752.1
insulin isoform X2 [Theropithecus gelada]	Theropithecus gelada	214	214	100%	1e-69	96.36%	147	XP_025211011.1
insulin [Tupaia chinensis]	Tupaia chinensis	188	188	100%	9e-60	91.82%	110	XP_006141129.1

Figure 2: Result of BLASTP using human insulin as query

The outcome of BLASTP search displays 100 protein sequences which are similar with the query protein. From the results, 4 protein sequences are selected at random to perform protein sequence data analysis along with the human insulin (AAA59172.1). At the end, a total of five protein sequences are employed to carry out the multiple sequence alignment using ClustalOmega.

The five proteins that are chosen for multiple sequence alignment are:

1. AAA59172.1 insulin [Homo sapiens]
2. XP_024110665.1 insulin isoform X2 [Pongo abelii]
3. XP_008002825.1 insulin isoform X2 [Chlorocebus sabaeus]
4. XP_053418912.1 insulin [Nycticebus coucang]
5. XP_052029588.1 insulin [Apodemus sylvaticus]

The FASTA format of each of the chosen proteins are retrieved from NCBI website, manually compiled into a single FASTA file and saved as insulin.fasta for further use.

4.2 ClustalOmega Implementation

4.2.1 Importing Libraries

```
library(msa)
library(Biostrings)
library(seqinr)
library(ape)
```

```
> library(msa)
> library(Biostrings)
> library(seqinr)
> library(ape)
```

Figure 3: Code snippet for importing libraries

4.2.2 Importing Dataset

```
file_path <- file.path("D:", "UTM", "Course", "Y3S2", "Bioinformatics
Database", "insulin.fasta")
insulin <- readAAStringSet(file_path)
insulin
```

```
> file_path <- file.path("D:", "UTM", "Course", "Y3S2", "Bioinformatics Database", "insulin.fasta")
> insulin <- readAAStringSet(file_path)
> insulin
AAStringSet object of length 5:
    width seq                                                 names
[1]   110 MALWMRLLPLLALLALWGPDPAAA...QKRGIVEQCCTSICSLYQLENYCN AAA59172.1 insuli...
[2]   110 MALWMRLLPLLALLALWGPDPAQA...QKRGIVEQCCTSICSLYQLENYCN XP_024110665.1 in...
[3]   144 MGSETIKPVGAQQPSTLRDGCIRR...QKRGIVEQCCTSICSLYQLENYCN XP_008002825.1 in...
[4]   110 MALWMRLLPLLALLALWGPQPASA...QKRGIVEQCCTSICSLYQLENYCN XP_053418912.1 in...
[5]   110 MALWMRFLPLLVLLFLWEPNPAQA...QKRGIVDQCCTSICSLYQLENYCN XP_052029588.1 in...
```

Figure 4: Code snippet for importing dataset

4.2.3 ClustalOmega Alignment

```
myClustalOmegaAlignment <- msa(insulin, "ClustalOmega")
myClustalOmegaAlignment
```

```
> myClustalOmegaAlignment <- msa(insulin, "ClustalOmega")
using Gonnet
> myClustalOmegaAlignment
ClustalOmega 1.2.0

Call:
   msa(insulin, "ClustalOmega")

MsaAAMultipleAlignment with 5 rows and 144 columns
    aln                                                   names
[1] ----------------------------...AQQKRGIVDQCCTSICSLYQLENYCN XP_052029588.1 in...
[2] ----------------------------...SLQKRGIVEQCCTSICSLYQLENYCN AAA59172.1 insuli...
[3] ----------------------------...SLQKRGIVEQCCTSICSLYQLENYCN XP_024110665.1 in...
[4] MGSETIKPVGAQQPSTLRDGCIRRGQQ...SLQKRGIVEQCCTSICSLYQLENYCN XP_008002825.1 in...
[5] ----------------------------...PPQKRGIVEQCCTSICSLYQLENYCN XP_053418912.1 in...
Con ----------------------------...SLQKRGIVEQCCTSICSLYQLENYCN Consensus
```

Figure 5: Code snippet for summary of ClustalOmega alignment

```
print(myClustalOmegaAlignment, show = "complete")
```

```
> print(myClustalOmegaAlignment, show = "complete")

MsaAAMultipleAlignment with 5 rows and 144 columns
    aln (1..56)                                                names
[1] ---------------------------------MALWMRFLPLLVLLFLWEPNPA XP_052029588.1 in...
[2] ---------------------------------MALWMRLLPLLALLALWGPDPA AAA59172.1 insuli...
[3] ---------------------------------MALWMRLLPLLALLALWGPDPA XP_024110665.1 in...
[4] MGSETIKPVGAQQPSTLRDGCIRRGQQAGHCPSAMALWMRLLPLLALLALWGPDPV XP_008002825.1 in...
[5] ---------------------------------MALWMRLLPLLALLALWGPQPA XP_053418912.1 in...
Con ---------------------------------MALWMRLLPLLALLALWGPDPA Consensus

    aln (57..112)                                              names
[1] QAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQ XP_052029588.1 in...
[2] AAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQ AAA59172.1 insuli...
[3] QAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQ XP_024110665.1 in...
[4] PAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQ XP_008002825.1 in...
[5] SAFVNQHLCGSHLVEALYLVCGERGFFYTPKARRDMEDPQVGQVGLGGSPITGDLQ XP_053418912.1 in...
Con ?AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQ Consensus

    aln (113..144)                  names
[1] TLALEVAQQKRGIVDQCCTSICSLYQLENYCN XP_052029588.1 in...
[2] PLALEGSLQKRGIVEQCCTSICSLYQLENYCN AAA59172.1 insuli...
[3] PLALEGSLQKRGIVEQCCTSICSLYQLENYCN XP_024110665.1 in...
[4] PLALEGSLQKRGIVEQCCTSICSLYQLENYCN XP_008002825.1 in...
[5] PLALDVPPQKRGIVEQCCTSICSLYQLENYCN XP_053418912.1 in...
Con PLALEGSLQKRGIVEQCCTSICSLYQLENYCN Consensus
```

Figure 6: Code snippet for printing the complete result of ClustalOmega alignment

4.2.4 Distance Matrix

```
myClustalOmegaAlignment2 <- msaConvert(myClustalOmegaAlignment, type =
"seqinr::alignment")
dis <- dist.alignment(myClustalOmegaAlignment2, "identity")
as.matrix(dis)[1:5, "AAA59172.1 insulin [Homo sapiens]", drop = FALSE]
```

```
> myClustalOmegaAlignment2 <- msaConvert(myClustalOmegaAlignment, type = "seqinr::alignment")
> dis <- dist.alignment(myClustalOmegaAlignment2, "identity")
> as.matrix(dis)[1:5, "AAA59172.1 insulin [Homo sapiens]", drop = FALSE]
                                                    AAA59172.1 insulin [Homo sapiens]
XP_052029588.1 insulin [Apodemus sylvaticus]                                0.41560471
AAA59172.1 insulin [Homo sapiens]                                           0.00000000
XP_024110665.1 insulin isoform X2 [Pongo abelii]                            0.09534626
XP_008002825.1 insulin isoform X2 [Chlorocebus sabaeus]                     0.16514456
XP_053418912.1 insulin [Nycticebus coucang]                                 0.36927447
```

Figure 7: Code snippet for computing the distance matrix

4.2.5 Phylogenetic Tree

```
phyTree <- nj(dis)
plot(phyTree, main = "Phylogenetic Tree of Insulin Sequences")
```

```
> phyTree <- nj(dis)
> plot(phyTree, main = "Phylogenetic Tree of Insulin Sequences")
```

Figure 8: Code snippet for constructing phylogenetic tree

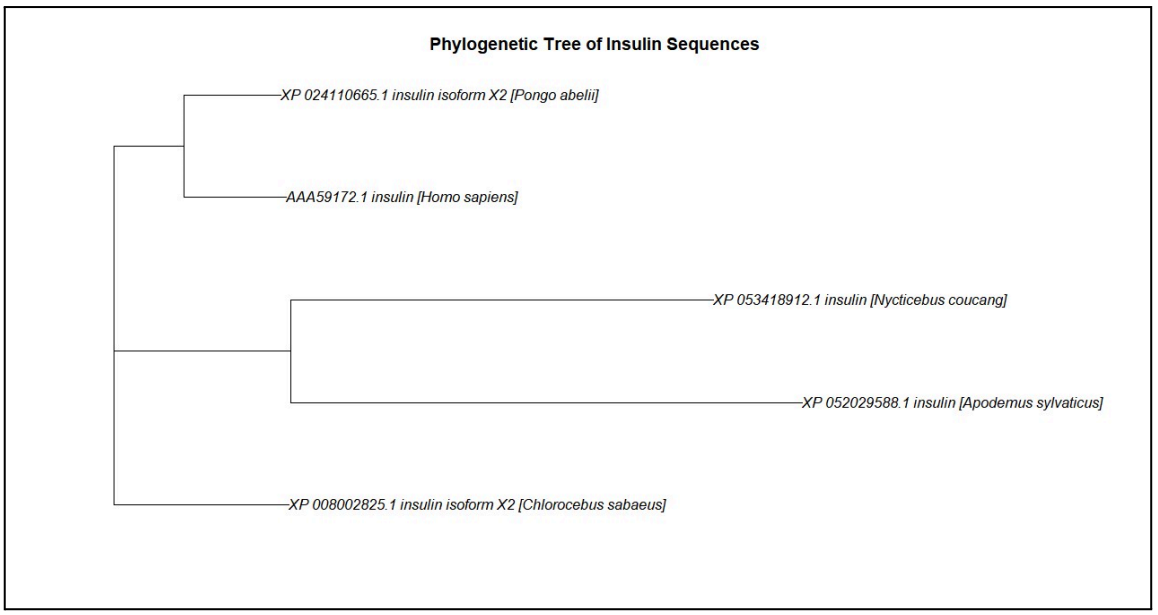


Figure 9: Phylogenetic tree of ClustalOmega alignment

5. Analysis and Discussion of Result

The sequence alignment of the 5 insulin proteins of different organisms is created by Clustal Omega as shown below.

```
MsaAAMultipleAlignment with 5 rows and 144 columns
      aln (1..43)                                      names
[1] ----------------------------------MALWMRFLP XP_052029588.1 in...
[2] ----------------------------------MALWMRLLP AAA59172.1 insuli...
[3] ----------------------------------MALWMRLLP XP_024110665.1 in...
[4] MGSETIKPVGAQQPSTLRDGCIRRGQQAGHCPSAMALWMRLLP XP_008002825.1 in...
[5] ----------------------------------MALWMRLLP XP_053418912.1 in...
Con ----------------------------------MALWMRLLP Consensus

      aln (44..86)                                     names
[1] LLVLLFLWEPNPAQAFVKQHLCGSHLVEALYLVCGERGFFYTP XP_052029588.1 in...
[2] LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTP AAA59172.1 insuli...
[3] LLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTP XP_024110665.1 in...
[4] LLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTP XP_008002825.1 in...
[5] LLALLALWGPQPASAFVNQHLCGSHLVEALYLVCGERGFFYTP XP_053418912.1 in...
Con LLALLALWGPDPA?AFVNQHLCGSHLVEALYLVCGERGFFYTP Consensus

      aln (87..129)                                    names
[1] MSRREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQC XP_052029588.1 in...
[2] KTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC AAA59172.1 insuli...
[3] KTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC XP_024110665.1 in...
[4] KTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC XP_008002825.1 in...
[5] KARRDMEDPQVGQVGLGGSPITGDLQPLALDVPPQKRGIVEQC XP_053418912.1 in...
Con KTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC Consensus

      aln (130..144)  names
[1] CTSICSLYQLENYCN XP_052029588.1 in...
[2] CTSICSLYQLENYCN AAA59172.1 insuli...
[3] CTSICSLYQLENYCN XP_024110665.1 in...
[4] CTSICSLYQLENYCN XP_008002825.1 in...
[5] CTSICSLYQLENYCN XP_053418912.1 in...
Con CTSICSLYQLENYCN Consensus
```

The length of insulin isoform X2 of Chlorocebus sabaeus is different from the other four insulin proteins which is 144 amino acids, whereas the other proteins are having length of 111 amino acids. In the alignment of five insulin proteins, the first block shows the alignment of the first 43 residues. The second block shows the alignment of residues 44 to 86, while the third block shows the alignment of residues 87 to 129 and the last block displays the alignment from residues 130 to 144.

From the first alignment to the 34th amino acid, the consensus sequences display as dash symbols which indicates that there is a gap. Then for the following residues, the five insulin proteins alignment are able to generate the consensus sequence except for residue 57th (highlighted with light green colour). The "?" symbol in the consensus sequence indicates that the algorithm is unable to find a majority of amino acid at that position among the five insulin proteins. In the other position, the algorithm is able to generate the consensus sequence because there are a majority of similar amino acids in those particular positions with only a minority of differences. The amino acids that are highlighted in red are the amino acids that are different from the amino acids of the other insulin proteins at the position.

.

	AAA59172.1 insulin [Homo sapiens]
XP_052029588.1 insulin [Apodemus sylvaticus]	0.41560471
AAA59172.1 insulin [Homo sapiens]	0.00000000
XP_024110665.1 insulin isoform X2 [Pongo abelii]	0.09534626
XP_008002825.1 insulin isoform X2 [Chlorocebus sabaeus]	0.16514456
XP_053418912.1 insulin [Nycticebus coucang]	0.36927447

The values above represent the square root of the pairwise distances, where 0 indicates that the two sequences are identical. From the result of alignment, if the alignment value is closer to 1, it indicates that there is a strong cluster structure that is well separated and distinct to each other whereas for the value that is closer to 0, it indicates that the clustering structure is not significantly different. Thus, insulin isoform X2 of Pongo abelii is the sequence that is most identical to the insulin of Homo sapiens. Next, the second identical insulin is the insulin isoform X2 of Chlorocebus sabaeus with a pairwise distance square root of 0.165, then followed by insulin of Nycticebus coucang with 0.369 and lastly insulin of Apodemus sylvaticus with 0.416.

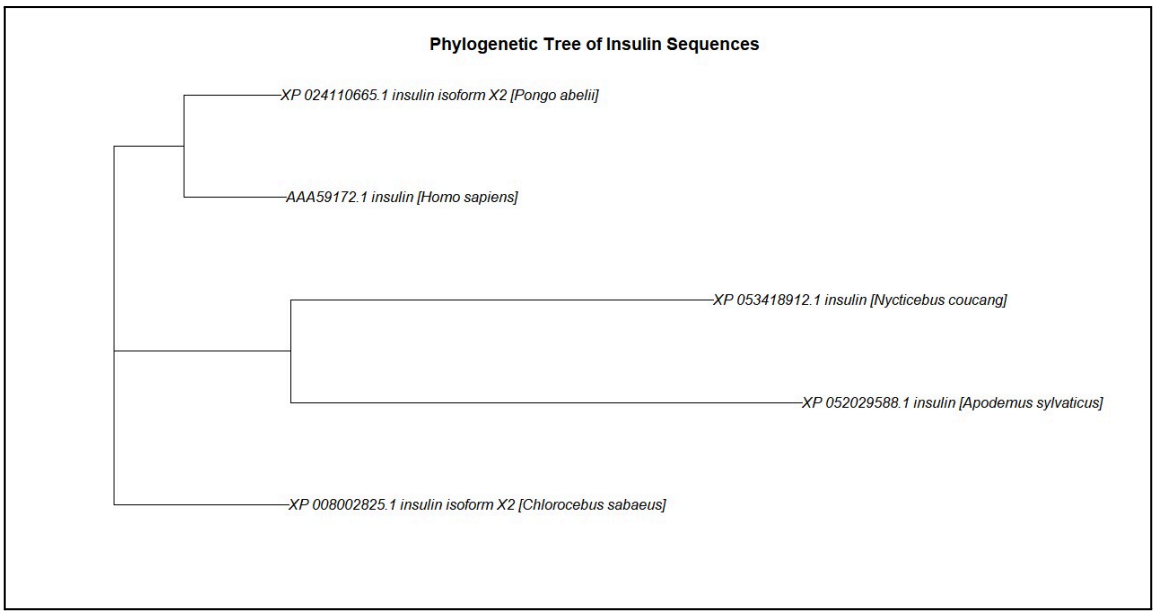


Figure 10: Phylogenetic tree of ClustalOmega alignment

The phylogenetic tree is created to visualise the evolutionary relationship between the five insulin sequences. It created a similar result with the pairwise distance matrix calculated in the previous part. According to the tree, insulin of Pongo abelii is evolutionary closer to the insulin of homo sapiens. Furthermore, the ancestor of the Pongo abelii insulin and Homo sapiens insulin is evolutionary closer to insulin isoform X2 of Chlorocebus sabaeus. On the other hand, the evolutionary relationship of Nycticebus coucang insulin is close to insulin of Apodemus sylvaticus, though both of them are evolutionary far with insulin of Pongo abelii, Homo sapiens and Chlorocebus sabaeus.

6. Conclusion

This project report consists of 6 chapters, including the (1) introduction, (2) literature review, (3) methodology, (4) implementation, (5) result analysis and discussion, as well as (6) conclusion. The first chapter includes a concise explanation of the key ideas behind the project. Literature review provides an overview on the NCBI database and the application of protein sequence in Multiple Sequence Alignment (MSA). The following chapter is methodology that maps out the operational framework of the project. The implementation of the project including the program written in R programming language is shown in Chapter 4 and the results are discussed in Chapter 5. Lastly, Chapter 6 includes the conclusion for the whole project.

The achievement of this project is the implementation of ClustalOmega to perform MSA to identify the relationship between the five insulin proteins. In short, according to the results of MSA, the insulin of Pongo abelii is the most similar protein to the Homo sapiens insulin. Throughout the project, we are able to realise the significant contribution of MSA to the bioinformatics field especially on data analysis of protein sequences. It provided an opportunity for us to have hands-on experiences on what we have learned in the lectures. Lastly, the project also helps us to have a deeper understanding of MSA.

7. References

EMBL-EBI UK. (n.d.). Pairwise Sequence Alignment Tools < EMBL-EBI. EMBL-EBI. https://www.ebi.ac.uk/Tools/psa/

Yashvardhini, N. and Kumar Jha, D. (2020) 'Occurrence of Recurrent Mutations in SARS-CoV-2 Genome and its Implications in the Drug Designing Strategies', *Journal of Pharmacy and Pharmacology Research*, 04(04), pp. 96–102. doi: 10.26502/fjppr.035.