1.

A).

* Big Data Analytics in Bioinformatics has advanced the way of finding drugs to treat certain diseases. For instance, researchers, computer scientists, bioinformaticians found similarities between the SARS-CoV-2 and SARS-CoV-1 genome sequences using Big data. They analyzed patterns in existing genome data available and which lead to the fast-tracked development of vaccines for Covid-19. Before the use of software and machine learning techniques in the drug development process, it took years of testing to develop a drug. The process of finding the effects of drugs on humans was all manual. Now a lot of this is reduced as scientists can find similar structures of known proteins in novel viruses that can help predict how the body will react to certain drugs.
* Big data analytics has helped in the study of the genomes of all organisms faster than ever before. Using analytical software and Big data we can now check the similarities between DNAs of different organisms. This has helped researchers to learn about the functioning of the same proteins in different organisms in a much faster way. Also, Big data analytics has helped researchers learn about how genes interact with other biomolecules.

B).

EMBL-EBI- https://www.ebi.ac.uk/

PDB- https://www.rcsb.org/pages/search\_features

NCBI/GenBank- https://www.ncbi.nlm.nih.gov/gene/

C).

NCBI- https://www.ncbi.nlm.nih.gov/structure/

PDB- https://www.rcsb.org/pages/search\_features

SWISS-MODEL Repository- https://swissmodel.expasy.org/repository/

D).

Genome is the collection of all the information which is inside our body's cells in its nucleus responsible for building all the proteins, cells and organs inside our body. A genome comprises DNA (Deoxyribonucleic acid) which is a double-strand helical structure present in 23 pairs of chromosomes. Each of the chromosomes has a copy of DNA from both mother and father. So, a genome is a set of instructions for the functions of all organisms.

USCS- https://genome.ucsc.edu/cgi-bin/hgGateway

Ensembl Genomes- http://ensemblgenomes.org/

NCBI- https://www.ncbi.nlm.nih.gov/genome/

E).

Machine learning is very important in Bioinformatics. As the number of DNA sequences discovered by science is increasing, machine learning is helping to predict the protein structures using DNA sequences. Protein structures are very informative of the functions of a protein which help scientist reduce the effort to learn about a gene from scratch. Many software like SWISS-MODEL repository uses machine learning techniques to create structures for unknown sequences. Machine learning is also used in predicting genetic disorders by analyzing genes. Softwares like DisEMBL can be very useful to predict diseases and provide early treatment of genetic disorders or in controlling symptoms before full development. There are multiple applications of machine learning in Bioinformatics, and many are still unexplored. Machine learning has evolved the research in Bioinformatics and has huge potential to solve many problems in this domain.

2.1

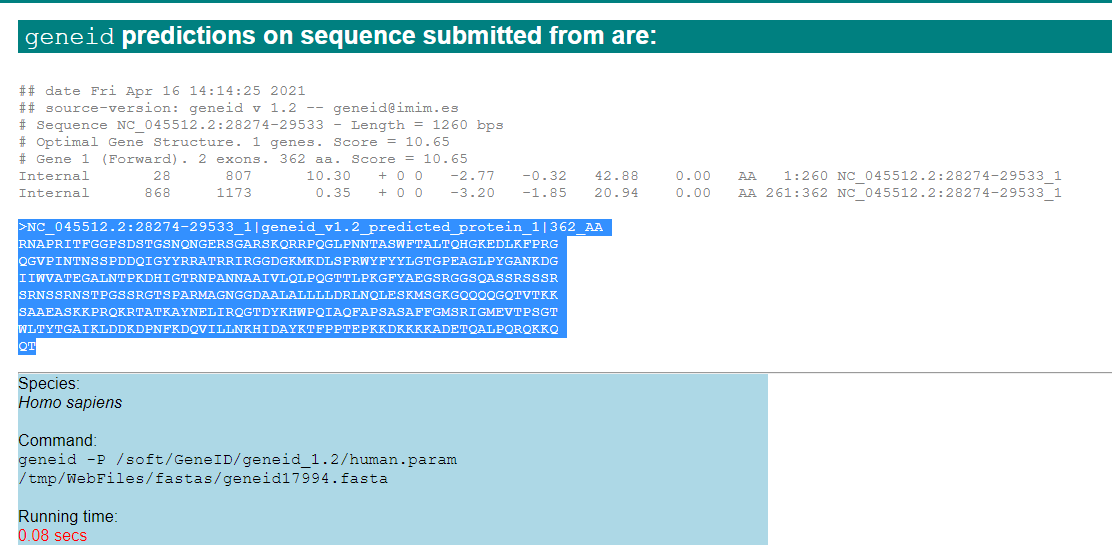
A). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a single strand RNA virus that causes coronavirus disease. Virus functions when encounters host cells and use Ribosome of the host to read the RNA and create proteins and supporting non-structural proteins essential for the virus. This results in creating all the necessary structures for a new copy of the virus. The virus once spread in the human body causes fever-like symptoms and within a period of 14 days. Symptoms vary from mild fever to acute lung infection.

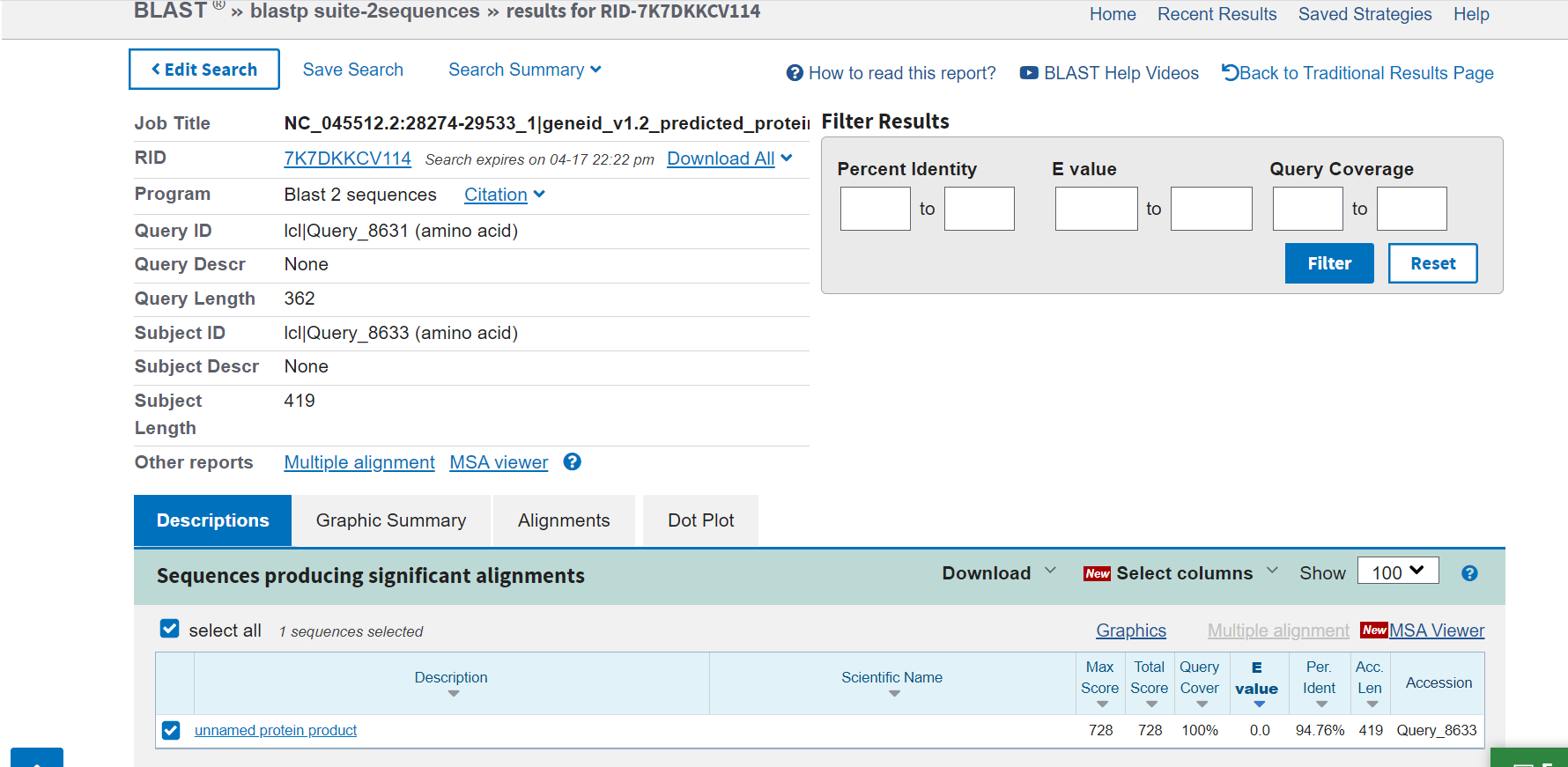
B). SARS-CoV-2 gene N is responsible for packing the RNA helical structure. It plays an important role in the assembly of virus and interactions with other viral genomes. It is also responsible for viral transcription and replication.

2.2

A). The predicted protein sequence by the geneid software is approximately 95% like the NCBI protein sequence. A few amino acids in the beginning and end were not predicted correctly in the protein sequence. This is because prediction is based on analysis of nucleotides showing similarity to a known protein. It may or may not be 100% accurate.

I used PROTEIN BLAST on NCBI to check the similarities in two protein sequences.





B). BLAST is a tool to search and predict similar proteins using nucleotide/protein sequences. It is a part of NCBI. We can also use the ClustalW tool by Protein Information Resource (PIR) a protein database to predict protein sequences.

2.3

A). Protein Name- *MERS CoV nucleocapsid-* N PROTEIN

B). Sequence Identity - 55.41

C). Template Name- 4ud1.1

D). Screen-shot of Model’s sequence alignment and template-

Graphical user interface, text, application

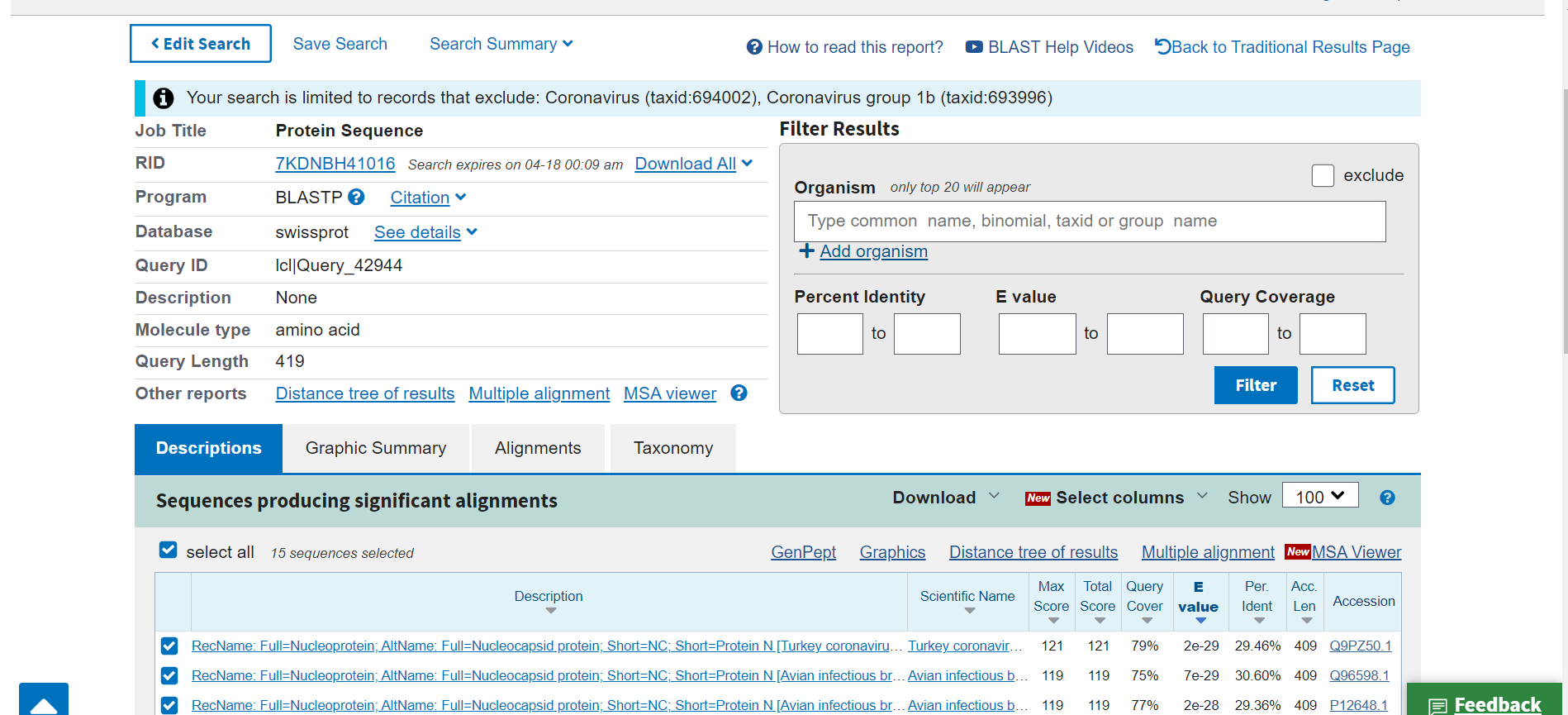
Description automatically generated

A picture containing graphical user interface

Description automatically generated

2.4

A).



The results show that the protein sequences producing significant alignments with sequence of 419 amino acids from question 2.3. The protein sequences aligned exclude Coronavirus (taxid:694002), Coronavirus group 1b (taxid:693996).

B).

The protein name is Protein N [Avian infectious bronchitis virus (strain DE072)]. Protein is 31.38 % similar to the protein sequence from question 2.3. The results show that the protein has 75 % query coverage. As the name suggests ‘Avian’: the protein sequence is related to infection in birds.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

C).

Generated a 3D model using the Swiss - Model:

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

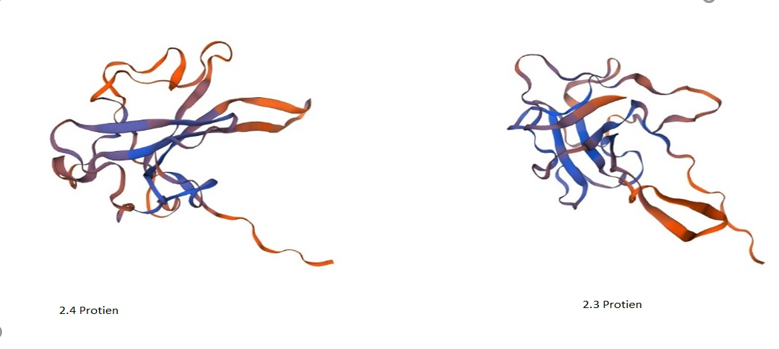
D).

Screen-shot of the results page:

Graphical user interface, application

Description automatically generated

E).



I have used the protein structure with the highest coverage. The MERS CoV nucleocapsid has the highest coverage that is 0.35. By observation, it looks that the protein structure generated in 2.3 has some similarities to the one produced for this protein sequence. Both structures use the same template 4ud1.1.A but have different coverage and sequence identities. However, there are many computational techniques like Dynamic Programing to find structural similarities.