SEQUENCE SIMILARITY SEARCH OF NSP4 PROTEIN (CORONA VIRUS) USING PSI-BLAST

A

Project Report Submitted To

BUXI JAGABANDU BIDYADHAR AUTONOMOUS COLLEGE BHUBANESWAR

By

DHRUTI RANJAN MOHANTY

Roll No: 2018MBI014

DEPARTMENT OF BIOINFORMATICS

BUXI JAGABANDHU BIDYDHAR AUTONOMOUS COLLEGE

BHUBANESWAR-751014

BJB (A) COLLEGE, BHUBANESWAR

DEPARTMENT OF BIOINFORMATICS

Mr. Sabyasachi Mohanty
H.O.D, Dept. of IMSc.BI

Date:

CERTIFICATE – I

This is to certify that the project report entitled **"Sequence Similarity Search of Nsp4 Protein (Corona Virus) Using Psi-Blast"** was submitted towards the fulfillment of Paper code: 604 of 6th Semester **Integrated MSc. Bioinformatics** of the **BJB (A) College, under Utkal University**, **Bhubaneswar** embodies a faithful record of bonafide and original research work carried out by **Dhruti Ranjan Mohanty (2018MBI014)** under my guidance and supervision. No part of this report has been submitted for any other degree or diploma.

The assistance and help received during the course of investigation has been fully acknowledged.

Place: Bhubaneswar
Date:

Mr.Sabyasachi Mohanty
H.O.D,Dept. of IMSc.BI
BJB(A) COLLEGE

BJB (A) COLLEGE, BHUBANESWAR

DEPARTMENT OF BIOINFORMATICS

Mr. Sabyasachi Mohanty
H.O.D, Dept. of IMSc.BI

Date:

CERTIFICATE – II

This is to certify that the thesis entitled **"Sequence Similarity Search of Nsp4 Protein (Corona Virus) Using Psi-Blast"** submitted by **Dhruti Ranjan Mohanty** to **BJB(A) College**, **Bhubaneswar** towards the partial fulfillment of 6th semester for the Bachelor Degree of Integrated MSc. Bioinformatics of the BJB (A) COLLEGE, Utkal University, is eligible for bonafied research work form oral examination on the same in the presence of External Examiner.

Internal Examiner
Date:

External Examiner
Date:

BJB (A) COLLEGE, BHUBANESWAR

DEPARTMENT OF BIOINFORMATICS

DECLARATION

I do hereby declare that I gave undertaken the project work entitled **"Sequence Similarity Search of Nsp4 Protein (Corona Virus) Using Psi-Blast"** submitted to BJB (A) COLLEGE, Bhubaneswar in partial fulfillment of Bachelor degree in Integrated MSc. Bioinformatics is an original work done by me under the supervision of **Mr. Sabyasachi Mohanty, HOD, Dept. of Bioinformatics, BJB Autonomous College, Bhubaneswar.** This thesis comprises of a bonafide research work and no part of this report has been submitted for any other degree or diploma.

Dhruti Ranjan Mohanty
2018MBI014

BJB (A) COLLEGE, BHUBANESWAR

DEPARTMENT OF BIOINFORMATICS

BIO-DATA

Name	:	Dhruti Ranjan Mohanty
Roll No.	:	2018MBI014
Title of Thesis	:	Sequence Similarity Search of Nsp4 Protein (Corona Virus) Using Psi-Blast
Semester for Which Report Submitted	:	6th Semester
Paper Code	:	604
Name of the Department	:	Department Of Bioinformatics
College & University	:	BJB (A) College, Utkal University, Bhubaneswar
Year of Submission	:	2021
Name of Advisor	:	Mr. Sabyasachi Mohanty

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to my guide **Mr. Sabyasachi Mohanty, H.O.D, Dept. of Bioinformatics, Buxi Jagabandhu Bidyadhar Autonomous College** for his exemplary, monitoring and constant encouragement throughout the period of this thesis. His blessing, help and guidance given by his time to time shall carry me a long way in the journey of life on which I am about to embark.

I express my courteous and considerate gratitude to my faculty members **Mr. RAKESH RANJAN OJHA (Lecturer, Dept. of Bioinformatics)** and **Miss. AMRITA RAY (Lecturer, Dept. of Bioinformatics)** for their kind support and precious suggestions during my research work. The work would have been incomplete without their help. I also take this opportunity to express a deep sense of gratitude to them.

I express my sincere thanks to all my classmates, my seniors and all the persons who have helped me directly or indirectly whose name could not find a separate place, for their deep concern, selfless help, fulfilled and friendly attitude which avoided me from solitudeness and nostalgia during my project days.

With sincere respect and obligation I bow down before my parents, elders and teachers for being there for me throughout my journey of science in all my good and bad times with their unbound love, constant support, immeasurable moral support and blessings to build my carrier. I also owe my thanks to the Almighty Lord Jagannath whose omnipresence blessings and touch of spiritualism helped me to build a nice human out of myself and giving a great strength to pass through all ordeals of life.

Department of Bioinformatics
BJB (A) College, Bhubaneswar

CONTENTS

INTRODUCTION

Corona virus disease is defined as illness caused by a novel corona virus called Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2), which was first identified amid an outbreak of respiratory illness cases in Wuhan City, China. World Health Organization (WHO) first learned of this new virus on 31st December 2019. The name "Corona-virus" refers to the crown-like projections on the pathogen's surface.

The emergence of SARS-CoV-2 was first observed when cases of unexplained pneumonia were noted in the city of Wuhan, China. During the first weeks of the epidemic in Wuhan, an association was noted between the early cases and the Wuhan Huanan Seafood Wholesale Market (hereafter referred to as the "Huanan market"); cases were mainly reported in operating dealers and vendors. The authorities closed the market on 1 January 2020 for environmental sanitation and disinfection. The market, which predominantly sold aquatic products and seafood as well as some farmed wild animal products, was initially suspected to be the epicenter of the epidemic, suggesting an event at the human animal interface. Retrospective investigations identified additional cases with onset of disease in December 2019, and not all the early cases reported an association with the Huanan Market. Although the role of civets as intermediate hosts in the outbreak of severe acute respiratory syndrome (SARS) in 2002-2004 had been favoured and a role for pangolins in the outbreak of COVID-19 was initially posited, subsequent epidemiological and epizootic studies have not substantiated the contribution of these animals in transmission to humans. The possible intermediate host of SARS-CoV2 remains elusive. Bats have been identified as the hosts of a series of important zoonotic viruses (for example, Nipah virus, Hendra virus and SARS-CoV), including coronaviruses with considerable genetic diversity. Of particular relevance with regard to COVID-19 are those coronaviruses that were found to be associated with the outbreaks in humans of SARS in 2002 and the Middle East respiratory syndrome (MERS) in 2013. The causative virus of COVID-19 was rapidly isolated from patients and sequenced, with the results from China subsequently being shared and published in January 2020. The findings showed that it was a positive-stranded RNA virus belonging to the Coronaviridae family (a subgroup B beta coronavirus) and was new to humans. In the early work, analysis of the genomic sequence of the new virus (SARS-CoV-2) showed high homology with that of the coronavirus that caused SARS in 2002-2004, namely SARS-CoV (another subgroup B beta coronavirus). Over the next year extensive work globally on sequences and phylogeny followed and the results have been shared internationally and stored through the GISAID platform. SARS-CoV-2 also shares a 96.2% homology with a sequence of a strain of coronavirus (RaTG13) previously identified by genetic sequencing from a horseshoe bat sample (Rhinolophus species) and to a lesser extent with a strain isolated from pangolins. The RaTG13 virus sequence is the closest known sequence to SARS-CoV-2.

How it spreads:

- The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces.
- We can be infected by breathing in the virus if we are within close proximity of someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose or mouth.
- People of all ages who experience fever and/or cough associated with difficulty breathing or shortness of breath, chest pain or pressure, or loss of speech or movement should seek medical care immediately. If possible, call your health care provider, hotline or health facility first, so you can be directed to the right clinic.

Symptoms:

The most common symptoms of COVID-19 are:

- Fever
- Dry cough
- Fatigue

Other symptoms that are less common and may affect some patients include:

- Loss of taste or smell,
- Nasal congestion,
- Conjunctivitis (also known as red eyes)
- Sore throat,
- Headache,
- Muscle or joint pain,
- Different types of skin rash,
- Nausea or vomiting,
- Diarrhea,
- Chills or dizziness.

Symptoms of severe COVID-19 disease include:

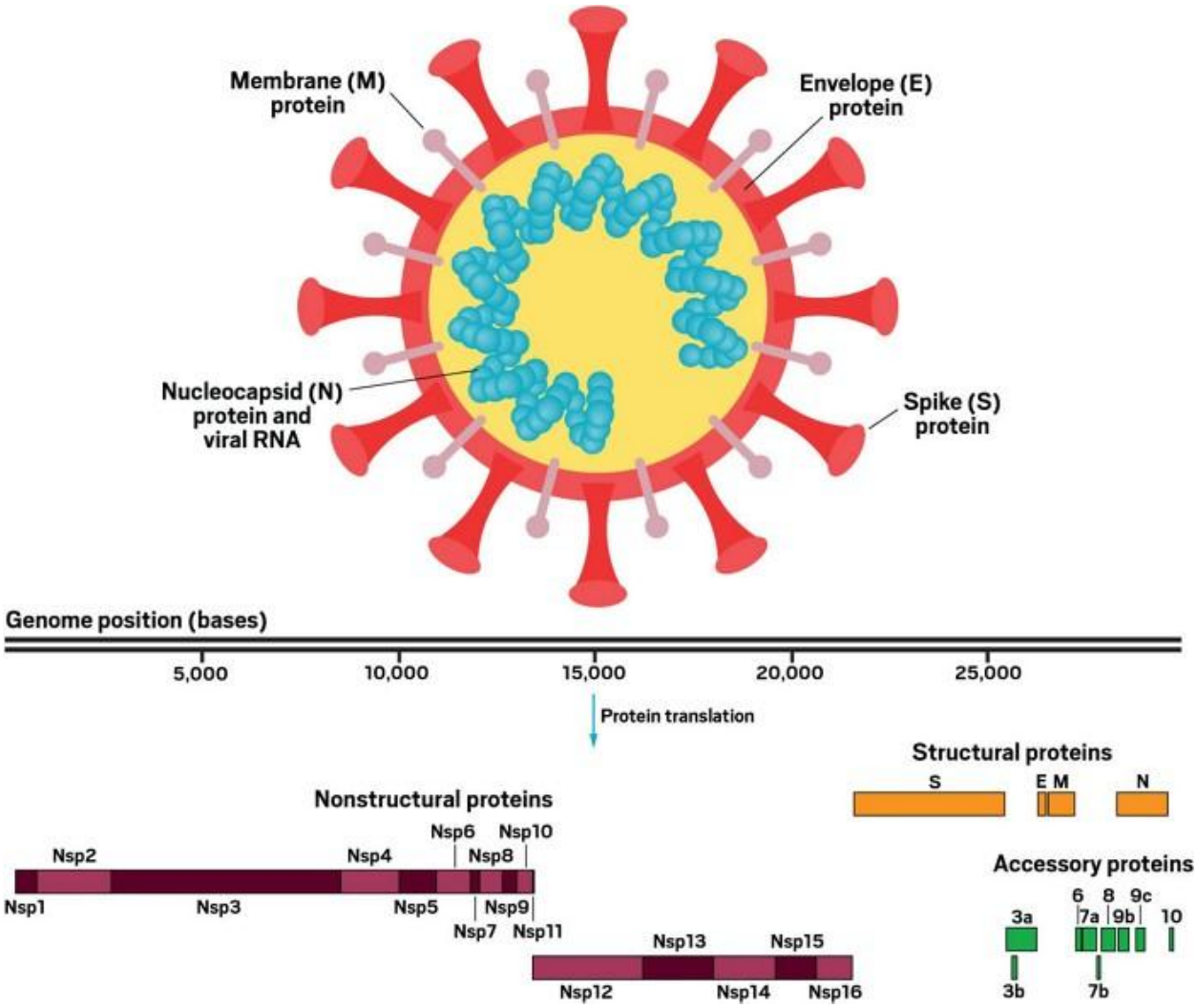
- Shortness of breath,
- Loss of appetite,
- Confusion,
- Persistent pain or pressure in the chest,
- High temperature (above 38 °C).

Other less common symptoms are:

- Irritability,
- Confusion,
- Reduced consciousness (sometimes associated with seizures)
- Anxiety,
- Depression,
- Sleep disorders

Covid Proteins:

The RNA genome of SARS-CoV-2 has 29,811 nucleotides, encoding for 29 proteins.



SARS-CoV-2 has four structural proteins (top): the E and M proteins, which form the viral envelope; the N protein (detail not shown), which binds to the virus's RNA genome; and the S protein, which binds to human receptors. The viral genome consists of more than 29,000 bases and encodes 29 proteins (bottom). The nonstructural proteins get expressed as two long polypeptides, the longer of which gets chopped up by the virus's main protease. This group of proteins includes the main protease (Nsp5) and RNA polymerase (Nsp12).

List of all Covid Protein:

Protein	Seq. Similarity (to SARS-CoV)	Function
Nsp1	91.1%	Suppress host antiviral response
Nsp2	82.9%	Unknown
Nsp3	86.5%	Viral replication
Nsp4	90.8%	Viral replication
Nsp5	98.7%	3c-Like Protease
Nsp6	94.8%	Viral replication
Nsp7	100.0%	Part of RNA polymerase
Nsp8	99.0%	Part of RNA polymerase
Nsp9	98.2%	Unknown
Nsp10	99.3%	Essential for Nsp16 methyltransferase activity
Nsp11	92.3%	Unknown
Nsp12	98.3%	RNA polymerase
Nsp13	100.0%	Helicase/triphosphatase
Nsp14	98.7%	3'-5' exonuclease
Nsp15	95.7%	Uridine-specific endoribonuclease
Nsp16	98.0%	RNA-cap methyltransferase

S	87.0%	Spike Protein Mediates binding to ACE2
Orf3a	85.1%	Activates the NLRP3 inflammasome
Orf3b	9.5%	Unknown
E	96.1%	Envelope Protein, Involved in virus morphogenesis and assembly
M	96.4%	Membrane glycoprotein, predominant component of the envelope
Orf6	85.7%	Type | IFN antagonist
Orf7a	90.2%	Unknown
Orf7b	84.1%	Unknown
Orf8	45.3%	Unknown
N	94.3%	Nucleocapsid phosphoprotein, binds to RNA genome
Orf9b	84.7%	Suppress host antiviral response
Orf9c	78.1%	Unknown
Orf10	--	Unknown

MATERIALS AND METHOD

The list of materials and process used is listed:

- Databases used:

1. National Center for Biotechnology Information (NCBI)
2. Non-redundant protein/nucleotide (nr/nt) database

DATABASES:

Databases are an organized collection of data in which one can submit and retrieve data easily. Curated database of small molecules includes interactions and functional effects of small molecules binding to their macromolecular targets. There are many type of small molecular database which can give information and structure and function of the taken proteins.

National Center for Biotechnology Information (NCBI):

- Website: http://www.ncbi.nlm.nih.gov/
- The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology.
- NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and many other kinds of biological data.
- Established as the government's response to the need for more and better information processing methods to deal with vast amount of data.
- A comprehensive website for biologists including :
  - Biology-related databases
  - Tools for viewing and analyzing
  - Automated systems for storing and retrieval
- NCBI has 3 collaborative databases:
  - GenBank
  - European Molecular Biology Laboratory (EMBL)
  - DNA Database of Japan (DDBJ)
- NCBI is now a leading source for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology.
- Today NCBI creates and maintains over 40 integrated databases for the medical and scientific communities as well as the general public.
- There are over 3 million visitors daily to its website, approximately 27 terabytes of data downloaded per day, and the number of users as well as downloads increases dramatically each year.

Non-redundant protein/nucleotide (nr/nt) database:

- The nr database is compiled by the NCBI (National Center for Biotechnology Information) as a protein database for Blast searches. It contains non-identical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF. The strengths of nr are that it is comprehensive and frequently updated.
- In that case, "nr/nt" stands for "non-redundant nucleotide." However, as you point out, NCBI also make separate databases available for download. In this case, "nr" is non-redundant protein; "nt" is non-redundant nucleotide.
- UniProtK /TrEMBL is 'non-redundant' in the sense that all identical, full-length protein sequences, provided they come from the same species, are represented in a single record. Fragments, isoforms, variants and so on, encoded by the same gene, are stored in separate entries.

PROTEIN INFORMATION (from NCBI):

Protein name : Nsp4

Accession ID : YP_009725300.1

GI : 1802476808

Molecular Weight : 56.2

Function : Nsp4 complex involved in viral replication

3D Structure :

- Tools used:

1. FASTA
2. BLAST

TOOLS:

Tools are used for implement, especially one held in the hand, used to carry out a particular function. We have categorized all the resources of databases as tools for the better understanding and a deeper knowledge about.

FASTA:

- FastA is a sequence comparison software that uses the method of Pearson and Lipman. The program compares a DNA sequence to a DNA database or a protein sequence to a protein database. Practically, FastA is a family of programs, which include: FastA, TFastA etc.
- In bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequence.
- How do we get a Fasta sequence?
  - Open NCBI website (http://www.ncbi.nlm.nih.gov/)
  - Select the Protein (from ALL databases) then, write the name of protein.
  - The list obtained; choice the specific protein and clicks on that.
  - Just below the name of the protein, FASTA is written, click on it.
  - You get new page having full information of protein sequence.

BLAST:

- The BLAST programs perform sequence-similarity searches against a variety of nucleotide and protein databases, returning a set of gapped alignments with links to full sequence records and to related transcript clusters (UniGene), 3D structures (MMDB).
- Blast is a program which uses specific scoring matrices (like PAM or BLOSSUM) for performing sequence-similarity searches against a variety of sequence databases, to give us high-scoring ungapped segments among related sequences.
- The BLAST algorithm is fast, accurate and web-accessible.
- It is relatively faster than other sequence similarity search tools.
- BLAST is a heuristic that finds short matches between two sequences and attempts to start alignments from these 'hot spots'.
- In addition to performing alignments, BLAST provides statistical information to help identifying the biological significance of the alignment; this is the 'expect' value, or false-positive rate.

- The method does not take the entire sequence space into account. After initial match, BLAST attempts to start local alignments from these initial matches.
- BLAST uses statistical theory to produce a bit score and expect value (E-value) for each alignment pair (query to hit). The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment.
- The lower the E-value, or the closer it is to zero, the more "significant" the match is.

Why do we use BLAST?

- It is a powerful tool used to search a database of DNA or protein sequences in order to find 'hits' that are similar to a query sequence.
- It is a computer algorithm that is available for use online at the National Center for Biotechnology Information (NCBI) website, as well as many other sites. It can rapidly align and compare a query DNA sequence with a database of sequences, which makes it a critical tool in ongoing genomic research.

PSI-Blast (Position Specific Iterated BLAST) :

- PSI-BLAST enhances the BLAST database searching method to incorporated PSSMs.
- Carefully constructed PSSM can find many distant members of a protein sequence family not easily found by a standard sequence search.
- Involves series of repeated steps or iterations.
- Perform standard BLAST search using a substitution matrix with a single query sequence.
- Obtain initial set of related sequences whose BLAST score gives an E-value smaller than a predetermined cut-off.
- Create a PSSM from alignments of these significant matches with the query sequence
- Scan PSSM against the database using a variant of the BLAST program to identify new sequences with suitably small E-values.
- If this second search finds some newly identified related sequences, use them to update the PSSM.

FASTA Format of Nsp4 protein:

```
>YP_009725300.1 nsp4 [Severe acute respiratory syndrome coronavirus 2]
KIVNNWLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFANKHAD
FDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTILRTTNGDFLHFLPRVFSAVGNICYTPSKLI
EYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVAYESLRPDTRYVLMDGSIIQFPNTYLEGSV
RVVTTFDSEYCRHGTCERSEAGVCVSTSGRWVLNNDYYRSLPGVFCGVDAVNLLTNMFTPLIQPIGALDI
SASIVAGGIVAIVVTCLAYYFMRFRRAFGEYSHVVAFNTLLFLMSFTVLCLTPVYSFLPGVYSVIYLYLT
FYLTNDVSFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFEEAALCTF
LLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHLAKALNDFSNSGSDVLYQPP
QTSITSAVLQ
```

Methods to perform PSI-Blast :

The PSI-blast is used to find distantly related proteins, what we need is the results in the final iteration.
The PSI-blast runs in the following five steps:

Step: 1

Select a query sequence and search it against a sequence database

Step: 2

PSI-BLAST constructs a multiple sequence alignment (of the sequences found) then creates a "profile" or specialized position-specific scoring matrix (PSSM).

Step: 3

The PSSM is used as a query against the database

Step: 4

PSI-BLAST estimates statistical significance (E values) of the sequences found.

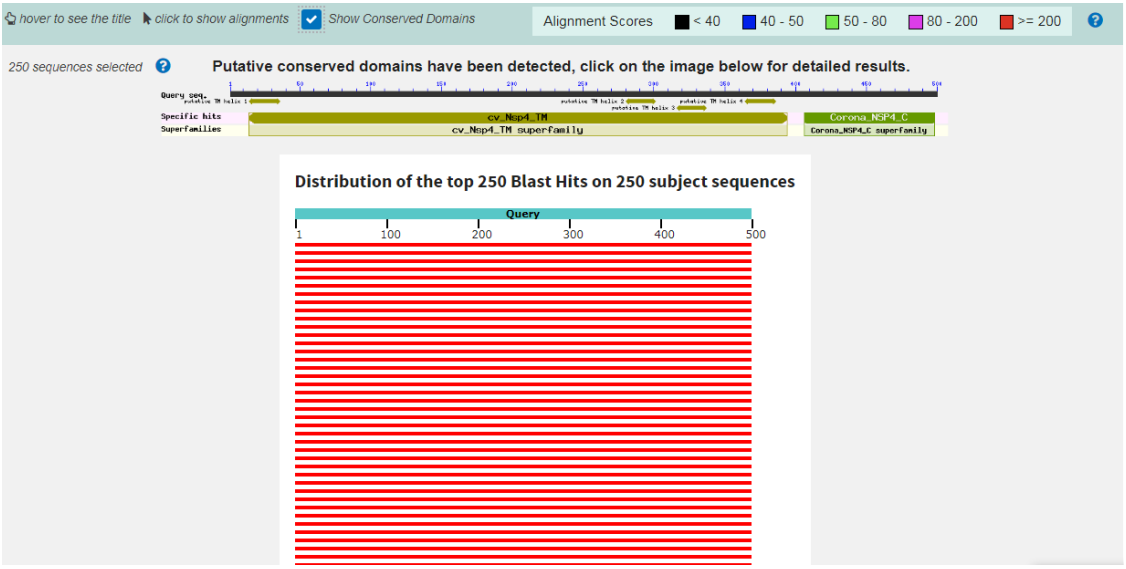
Step: 5

Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new PSSM is used as the query.

Different iterations use different PSSMs. A PSSM is constructed using sequences found in the last iteration by default. So the results are different in these iterations. Some previous sequences disappear in the latter iterations.
The reasons might be the following:
1. The query in each iteration (after the first iteration) is PSSM. We could see the PSSM as a sequence pattern. And the pattern changes after one iteration.
2. Because the PSI-blast is used to find distantly related sequences, so the PSI-blast should find the most likely sequence pattern, and uses it to find the sequences you want. These sequences are shown in the final iterations.
3. The disappeared sequences might not follow the best sequence pattern. And they are likely the spurious sequences.

RESULTS & DISCUSSION

1. Graphical representation of similar sequences:



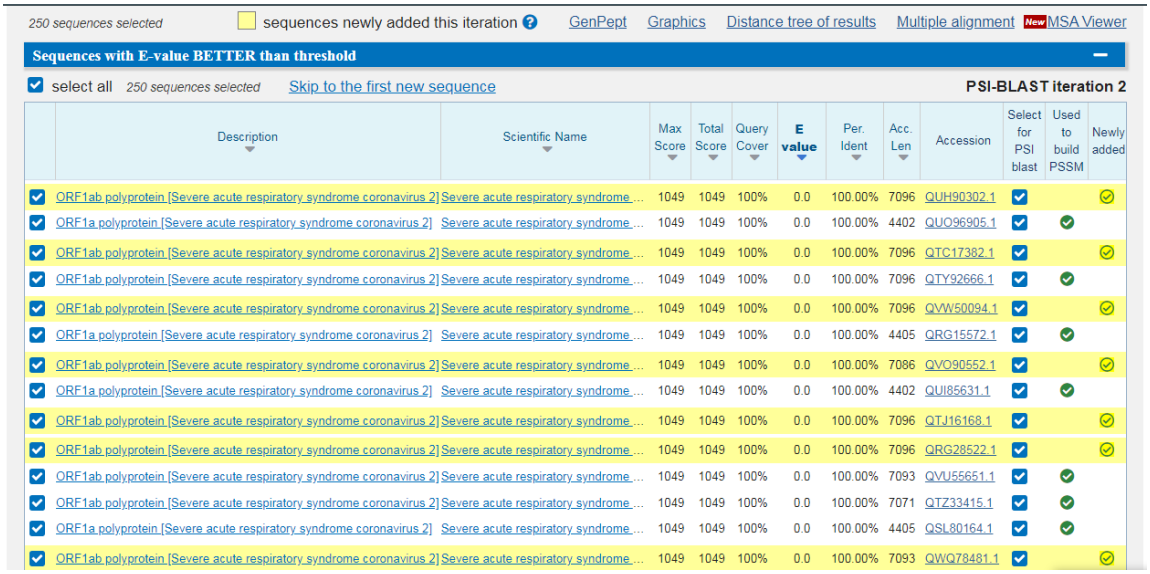
2. PSI-Blast iteration 1: (according to BLOSUM62)

Sequences with E-value BETTER than threshold

select all *250 sequences selected* **PSI-BLAST iteration 1**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1045	1045	100%	0.0	100.00%	4091	QRX46677.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1045	1045	100%	0.0	100.00%	4091	QRW61696.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1045	1045	100%	0.0	100.00%	4405	QTU35365.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QSY44566.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4377	QVI54549.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4402	QUB28452.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QSH79024.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4402	QWS79575.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QVP05639.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QSN74131.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QTD15899.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4405	QVO85985.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1044	1044	100%	0.0	100.00%	4402	QWA56959.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1043	1043	100%	0.0	100.00%	4402	QUP99458.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1043	1043	100%	0.0	100.00%	4402	QVG17046.1	☑		
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome ...	1043	1043	100%	0.0	100.00%	4156	QWE70294.1	☑		

3. PSI-Blast iteration 2: (according to PSSM)



250 sequences selected sequences newly added this iteration GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Sequences with E-value BETTER than threshold

select all 250 sequences selected Skip to the first new sequence **PSI-BLAST iteration 2**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QUH90302.1	✓		✓
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	4402	QUO96905.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QTC17382.1	✓		✓
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QTY92666.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QVW50094.1	✓		✓
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	4405	QRG15572.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7086	QVO90552.1	✓		✓
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	4402	QUI85631.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QTJ16168.1	✓		✓
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7096	QRG28522.1	✓		✓
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7093	QVU55651.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7071	QTZ33415.1	✓	✓	
ORF1a polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	4405	QSL80164.1	✓	✓	
ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome...	1049	1049	100%	0.0	100.00%	7093	QWQ78481.1	✓		✓

4. In PSI-Blast iteration 3 there is no new sequences were found above the threshold value in the database.

We notice first of all a bunch of very strong E-values better than threshold. We also notice that some protein sequences are highlighted in yellow and the sequences are there in white without the yellow highlighting are protein sequences that were found in the first iteration.

Any new sequence that was only found in the second iteration is highlighted in yellow and we see some new sequences with very strong evaluation that because the nature of the 2nd iteration is very different from the nature of the first iteration.

In the first iteration we started with a Nsp4 protein sequence; when we ran that first iteration we see that anything with an E-value of less than one times ten to the minus fourth (0.0001) is a homolog so now we see E-value use much less than one times ten to the minus fourth (0.0001) but let's remember in the second iteration we're not running the query sequence from Nsp4. we are now running a portion that represents both the Nsp4 sequence and a bunch of other ORF1a - ab polyprotein.

CONCLUSION

From the above analysis of similarity search it is concluded that, the E-value we see that the portion was mapped very well, it is suggested; it's probably a member of that protein family however in some cases a portion might not quite accurately represent a protein family may be because a non-relative got into that mix and it's usually best to treat these new yellow sequences as putative homologous meaning that they should be further checked for homology and we notice in the case that almost all of the sequence is highlighted in yellow says ORF1a - ab polyprotein which tells that they are probably indeed members of that protein family.

So we can see the PSI-Blast is a powerful tool for finding potential homologous of protein or DNA sequences that might not be found by the 1st iteration of blast.

Blast is not 100% sensitive, so it will not necessarily find all homologous in an initial database search.

PSI-Blast is a very powerful way to identify potential homologous that may have been missed by the first round of a blast search.

REFERENCES

1. *Whorld Health Organization,WHO*

2. Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402

3. Russo, E. and Bunk, S. (1999) Hot papers in bioinformatics, *Scientist* 13,15

4. Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST- a tool for discovery in protein databases. *Trends Biochem. Sci* 23, 444-447

5. Salamov, A.A et al. (1999) Combining sensitive database searches with multiple intermediates to detect distance homologues. *Protein Eng.*12, 95-100

6. Muller, A. *et al.*(1999) Benchmarking PSI-BLAST in genome annotation. *J.Mol.Biol.*293, 1257-1271

7. BIOINFORMATICS Principles and Applications by *ZHUMUR GHOSH* and *BIBEKANANDA MALLICK*

8. Schaffer, A.A. et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*29, 2994-3005

9. (http://cen.acs.org/biological-chemistry/infectious-disease/know-novel-coronaviruses-29-proteins/98/web/2020/04)

10. Article on *A SARS-CoV-2 protein interaction map reveals targets for drug repurposing*