CSC 314**, BLAST and Conserved Domain Identification**

**Introduction**

*The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares a nucleotide or protein* query *sequence to sequences in sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. [Modified from BLAST homepage,* [*https://blast.ncbi.nlm.nih.gov/Blast.cgi*](https://blast.ncbi.nlm.nih.gov/Blast.cgi)*]*

BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) works as follows:

1. All sequences are identified that contain exact matches of a fixed word size (for protein sequences, similarity is taken into account).
2. The alignment is then extended in both directions, using local alignment (the dynamic programming method).

Several versions of BLAST are available:

* *blastp* takes a protein sequence and BLASTs it against a protein database
* *blastn* takes a nucleotide sequence and BLASTs it against a nucleotide database
* *blastx* takes a nucleotide sequence, translates it, and BLASTs it against a protein database
* *tblastn* takes a protein sequence and BLASTs it against a translated nucleotide database

BLAST will return top scoring matches, which are high-scoring alignments with the query sequence; for each match, BLAST reports the following:

* Query cover: the percentage of the query sequence covered in the alignment
* E-value: the number of alignments that are expected by chance with the given score or better (this depends on the size of the query sequence and the database being searched – lower is better; typically 0.01 is used as a threshold for detecting homologous sequences)
* Percent Identity: the percent of the alignment positions where there are exact matches
* Positives: the percent similarity, for proteins only

**Conserved Domain Identification**

* A conserved domain contains conserved sequence patterns associated with distinct and independent functional or structural units of a protein
* Identifying protein domains can give us insight into a protein’s function
* NCBI has a conserved domain database (CDD); domains are based on multiple sequence alignments
* A version of BLAST can be used to search this CDD database (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>)
* When *blastp* is used, CDD is automatically searched and the results are returned with the graphical summary

**Background on COVID-19**

COVID-19 is the disease caused by the coronavirus “severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2)”. COVID-19 symptoms include fever, cough, shortness of breath, and breathing difficulties. More severe cases can lead to pneumonia, severe acute respiratory syndrome, and death. The disease was first identified in Wuhan, China in December 2019. The virus has since infected >676 million people worldwide and has resulted in around 6.8 million deaths (as of March 2023)[[1]](#footnote-1). In the United States, there have been over 103 million cases and over 1.1 million deaths[[2]](#footnote-2). Prior to the availability of vaccinations, the overall mortality rate was around 3% but is very dependent on age and pre-existing conditions, with mortality rates estimated at >10% for those 85 or older[[3]](#footnote-3). Estimates of R0, the number of new infections arising from a single infected individual, range from 1.5 – 3.5 (estimates are prior to vaccinations, mask wearing, and other restrictions that have been put in place)[[4]](#footnote-4).

**Using BLAST to understand COVID-19**

Since the genome of SARS-Cov-2 was published in early January 2020, researchers around the world have been analyzing the genome in order to better understand the virus, and to develop diagnostic tests, vaccines, and treatments for COVID-19. For an excellent summary about the SARS-Cov-2 genome, see the following: <https://www.nytimes.com/interactive/2020/04/03/science/coronavirus-genome-bad-news-wrapped-in-protein.html>

Here we will use BLAST to learn about SARS-Cov-2 by identifying sequences similar to SARS-Cov-2, whose genomic sequence is available here: <https://www.ncbi.nlm.nih.gov/nuccore/NC_045512>

***Question #1: Where did SARS-COV-2 come from?***

We can try to understand the origins of SARS-COV-2 by finding genomic sequences similar to the SARS-COV-2 genome. While this does not tell us with certainty how SARS-COV-2 originated, it does provide “circumstantial evidence” of it.

1. Open the SARS-Cov-2 genome sequence, available at the link above, and click Run BLAST.
2. Under database, select Standard databases and *Nucleotide collection*. Under organism, select *Betacoronavirus* and then exclude *SARS (taxid:694009).* Click BLAST.

What virus is most similar to SARS-Cov-2 (based on the top scoring alignment)?

What is the query coverage?

What is the percent identity?

***Question #2:* What is the function of NSP7?**

Following sequencing, the function of the protein now known as NSP7 was not known. What is the function of NSP7, whose protein can be found here:

<https://www.ncbi.nlm.nih.gov/protein/YP_009725303.1>

Note: we will use the “reference proteins” database when using BLAST (this database is a little bit smaller, so results will be slightly faster). One way of inferring protein function is through the identification of protein domains (functional regions of a protein identified by a particular amino acid sequence pattern). Conserved Domain Identification is a separate analysis than BLAST, but protein BLASTs will automatically identify conserved domains (available under the *Graphic Summary* tab).

***Question #3: Can genetic differences explain the infectiousness of SARS-Cov-2?***

There are several reasons, genetically, why one virus might be more infectious than another (why the virus spreads more easily from person to person). For example, a more infectious virus may have greater affinity for entering host cells, or may survive on surfaces for a longer period of time. The answer to this question is not known, but genomic comparisons can provide clues.

SARS, which caused an outbreak in 2003, is not as infectious as SARS-Cov-2. What genetic differences may account for this?

Repeat the BLAST from Question #2 but set the organism to include only SARS (taxid:694009)” and exclude “SARS-CoV-2 (taxid:2697049)”.

Is it likely that the protein NSP7 is responsible for differences in infectiousness between SARS-Cov-2? Why or why not?

**Question #4:** Repeat the above analysis for the surface (spike) glycoprotein, coded by the “S” gene, whose sequence is available here: <https://www.ncbi.nlm.nih.gov/protein/1796318598>

**Question #5: What is the function of *NSP2***

According to the NY Times summary, NSP2 is a “mystery protein”. Can we infer the function of NSP2 from a BLAST search. The protein sequence of NSP2 is given here: <https://www.ncbi.nlm.nih.gov/protein/1802476806>

1. <https://coronavirus.jhu.edu/map.html> [↑](#footnote-ref-1)
2. <https://www.nytimes.com/interactive/2023/us/covid-cases.html> [↑](#footnote-ref-2)
3. <https://ourworldindata.org/mortality-risk-covid#case-fatality-rate-of-covid-19-by-age> [↑](#footnote-ref-3)
4. <https://sph.umich.edu/pursuit/2020posts/how-scientists-quantify-outbreaks.html> [↑](#footnote-ref-4)