
	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ  PRODUCTON
BLAST Sequence Analysis	Effective Date: 02/15//2021 Checked by A. Siclair	Revision #1.0 M. Guzie



BACKGROUND

Sequence analysis is the bioinformatic process of using analytical tools and programs to study the comparative composition, structure, and function of biological macromolecules; primarily DNA, RNA, and proteins (polypeptides). Once a macromolecule has been sequenced, the sequence can be put into a program which will identify characteristics of the molecule based on comparison with other available sequences in a database. There are a variety of bioinformatic programs accessible to scientists today to analyze sequence properties in different ways, such as determining sequence similarities (BLAST), predicting how proteins will interact with DNA (ChIP-Seq), predicting how proteins will interact with other proteins (STRING) or with specific molecules/ligands (SwissDock), or giving a general molecular simulation to predict the behavior of molecules (OpenStructure).

The first techniques to sequence large biological molecules were developed in the mid-20th century. Renowned scientist Frederick Sanger greatly contributed to the field with his revolutionary chain termination sequencing technology developed in 1977; known as Sanger Sequencing, this technique involves the use of fluorescent dideoxynucleotides which act as terminators in the polymerization of DNA, as DNA polymerase needs the 3' OH to build upon. This creates many gene fragments which can then be ran on a gel and allow for sequencing based on length, revealing the gene's sequence. Methods of sequence analysis began to take off in the 1970's after sequencing techniques became more popularized, allowing more scientists to readily sequence DNA and then use analytical techniques to discover the functions of genes and molecules (Chain and Heather, 2016).

Techniques of sequence analysis allow scientists to compare and align sequences to see their similarity and predict their function, observe mutations, and study evolution. Sequence analysis can also be used to predict what the structure of a protein will be from the gene which encodes it and identify important reactive areas in the protein: active sites, ligand docks, and modification sites. This module covers the use of one of the most common tools of sequence analysis, BLAST (basic local alignment search tool). BLAST can be used to compare nucleotide (DNA/RNA) or protein sequence similarity to other sequences in a database. BLAST is useful to identify similar sequences which could possess similar functions or be homologs when new genes are discovered; for this reason, it is also applicable in the study of evolution. BLAST identifies similar molecular sequences by identifying short matching segments between the two sequences, then making local alignments. The BLAST algorithm was published in *The Journal of Molecular Biology* in 1990 and the program has grown in its use since (Lobo, 2008).

1. PURPOSE

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ  PRODUCTON
BLAST Sequence Analysis	Effective Date: 02/15//2021 Checked by A. Siclair	Revision #1.0 M. Guzie

The purpose of this procedure is to recognize the multitude of applicable uses of sequence analysis and to familiarize oneself with one of the most common applications, BLAST, by picking a gene of interest and using the database.

2. SCOPE



This procedure applies to qualified skills center users.

3. RESPONSIBILITY

- 3.1. It is the responsibility of the user to understand and perform the procedure described in this document.
- 3.2. It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure (in a computer lab, document your findings using the database.)
- 3.3. It is the responsibility of the user to become trained in the use of this application.

4. DEFINITIONS

- 4.1. BLAST: Basic Local Alignment Search Tool, a bioinformatics online database to compare macromolecule sequence similarities.
- 4.2. Nucleotide: The monomer building blocks which make up the nucleic acid polymers of DNA and RNA.
- 4.3. Amino Acids: The monomer building blocks which make up the polypeptide chain polymers which form proteins.
- 4.4. Query Sequence: The search sequence which is entered into the BLAST engine.
- 4.5. Max Score Value: Indicates similarity between best matching part of target sequence and query sequence.
- 4.6. Total Score Value: The overall sum of the BLAST scores from each segment which closely aligns between the match and the query segment; the total score value may be larger than the max score if the sequences align in multiple places.
- 4.7. Query Cover: The percentage of the alignment between the query sequence to the match sequence.
- 4.8. The Expect (E) Value: Indicates how likely it is that the match is due to chance; a higher value indicating the higher likelihood the match is due to chance, a lower value indicating the match to be significant. The cut off for an E-value will depend on several factors, such as the length of the query sequence and the size of the database.
- 4.9. Percent Identity: The percentage of identical bases in the best matching alignment region.

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ  PRODUCTON
BLAST Sequence Analysis	Effective Date: 02/15//2021 Checked by A. Siclair	Revision #1.0 M. Guzie

- 4.10. Homolog: A gene similar in structure and evolutionary origin to a gene in another species.
- 4.11. Contigs: In sequencing, overlapping DNA segments that are used to identify similar genes.

5. MATERIALS/EQUIPMENT

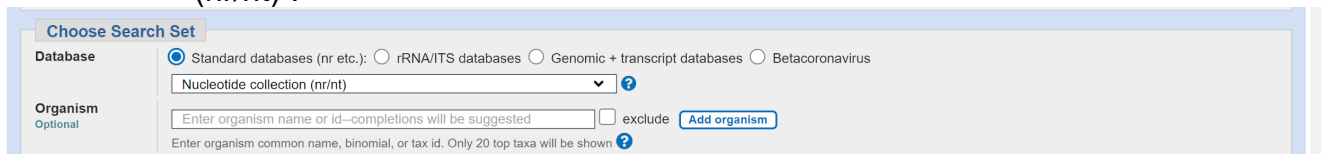
- 5.1. BLAST Database
- 5.2. National Institute of Health (NIH) Website
- 5.3. Method to record results (computer file or notebook.)

6. PROCEDURE



In this module you will perform a BLAST on both a DNA sequence and a polypeptide sequence. You can choose separate genes for each but it may be most beneficial and insightful to choose the same gene for both methods.

6.1. Nucleotide BLAST

- 6.1.1 Begin with a primary test by running a BLAST for the human myoglobin gene. The sequence of the myoglobin gene can be found here: <https://www.ncbi.nlm.nih.gov/nuccore/AH002877.2> (Scroll to the bottom where it says ORIGIN)
- 6.1.2. Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi> to perform the BLAST search for human myoglobin. Select Nucleotide BLAST.
- 6.1.3. Paste the gene sequence into the “Query Sequence” box.
- 6.1.4. Make sure the database is set to Standard Databases and “Nucleotide collection (nr/nt)”.



- 6.1.5. Under program selection, there are options to search for highly similar sequences (**megablast**), somewhat similar sequences (**blastn**), or more dissimilar sequences (**discontinuous megablast**). Megablast will search for the best matching sequences in the database, while blastn will result in a larger variety of sequences, including more from other organisms. Conduct each of these to observe the differences. Start with selecting megablast to find the sequences with the highest matches – select BLAST.
- 6.1.6. Megablast: Yielding the highest matches. Genes that match very well with the query gene will all appear beneath the query gene. Some of the matches might

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ  PRODUCTON
BLAST Sequence Analysis	Effective Date: 02/15//2021 Checked by A. Siclair	Revision #1.0 M. Guzie



just be segments of the query gene, or alternative mRNA transcripts due to splicing. Under the “descriptions” tab, each match will have several statistical values to the right (refer to DEFINITIONS section.)

Descriptions		Graphic Summary	Alignments	Taxonomy				
Sequences producing significant alignments								
<input checked="" type="checkbox"/> select all 100 sequences selected		Download <input type="button" value="New"/> Select columns <input type="button" value="Show"/> 100						
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Homo sapiens myoglobin (MB) gene, complete cds	Homo sapiens	4933	12936	99%	0.0	100.00%	6889	AH002877.2
<input checked="" type="checkbox"/> Homo sapiens DNA, chromosome 22, nearly complete genome	Homo sapiens	4922	62372	99%	0.0	99.93%	46684173	AP023482.1
<input checked="" type="checkbox"/> Homo sapiens myoglobin transcript variant 1 (MB) gene, complete cds	Homo sapiens	4911	12976	99%	0.0	99.85%	14045	DQ003030.1
<input checked="" type="checkbox"/> Homo sapiens isolate CHM13 chromosome 22	Homo sapiens	4872	69184	99%	0.0	99.59%	51353906	CP068256.1
<input checked="" type="checkbox"/> Eukaryotic synthetic construct chromosome 22	eukaryotic synt...	4867	47153	99%	0.0	99.55%	35194566	CP034501.1
<input checked="" type="checkbox"/> Homo sapiens myoglobin (MB), RefSeqGene on chromosome 22	Homo sapiens	4867	12775	99%	0.0	99.55%	23591	NG_007075.1
<input checked="" type="checkbox"/> Human DNA sequence from clone C1TF22-62D4 on chromosome 22, complete sequence	Homo sapiens	4867	10078	77%	0.0	99.55%	40665	AL049747.1
<input checked="" type="checkbox"/> Human myoglobin gene (exon 1) (and joined CDS)	Homo sapiens	4682	4682	38%	0.0	99.80%	3768	X00371.1

There are several different tabs that can be used to view results in different manners: shown here is “Descriptions”, showing matches and corresponding statistical significance, but there is also “Graphic Summary” which gives a visual representation of match alignments, “Alignments” which shows the actual nucleotide sequence of the alignments, and “Taxonomy” which shows relatedness and matches to other organisms. Explore these tabs to see how the information can be presented in different manners.

6.1.7. **Blastn:** Yielding somewhat similar sequences with more variability. After analyzing the results from the megablast, now perform a blastn. Return to the BLAST homepage and enter the Myoglobin nucleotide sequence again, this time selecting somewhat similar sequences “Blastn.” While many of the top results may be like the first megablast, as you scroll down you will find genes with lower percent identities and lower overall scores, indicating variations in the Myoglobin gene (or other genes) in other organisms.

<input checked="" type="checkbox"/> Canis lupus familiaris breed Labrador retriever chromosome 10a	Canis lupus familiaris	744	4181	57%	0.0	76.19%	69938001	CP050591.1
<input checked="" type="checkbox"/> Canis lupus familiaris breed Labrador retriever chromosome 10b	Canis lupus familiaris	741	4171	57%	0.0	76.15%	69942321	CP050611.1
<input checked="" type="checkbox"/> PREDICTED: Sus scrofa uncharacterized LOC110260651 (LOC110260651), ncRNA	Sus scrofa	736	850	25%	0.0	71.58%	1730	XR_002343916.1
<input checked="" type="checkbox"/> PREDICTED: Aotus nancymaae myoglobin (MB), transcript variant X2, mRNA	Aotus nancymaae	718	1203	15%	0.0	83.19%	1097	XM_012434301.1
<input checked="" type="checkbox"/> PREDICTED: Aotus nancymaae myoglobin (MB), transcript variant X1, mRNA	Aotus nancymaae	718	1492	18%	0.0	83.19%	1204	XM_012434300.1
<input checked="" type="checkbox"/> PREDICTED: Macaca mulatta uncharacterized LOC114670385 (LOC114670385), ncRNA	Macaca mulatta	669	732	7%	0.0	92.36%	2108	XR_003719884.1
<input checked="" type="checkbox"/> Homo sapiens MB full length open reading frame (ORF) cDNA clone (cDNA clone C22ORE...)	Homo sapiens	571	1240	10%	3e-157	99.38%	680	CR456516.1
<input checked="" type="checkbox"/> PREDICTED: Microcebus murinus myoglobin (MB), mRNA	Microcebus murinus	563	1029	16%	4e-155	78.62%	1132	XM_012787097.2
<input checked="" type="checkbox"/> Halichoerus grypus myoglobin gene fragment encoding first exon (and joined CDS)	Halichoerus grypus	530	530	8%	9e-145	80.79%	603	V00471.1
<input checked="" type="checkbox"/> Oryzias latipes strain HSOK chromosome 18	Oryzias latipes	464	3964	9%	3e-125	75.74%	30416519	CP020638.1
<input checked="" type="checkbox"/> Oryzias latipes strain HNI chromosome 23	Oryzias latipes	460	5068	9%	1e-123	76.36%	22812639	CP020801.1
<input checked="" type="checkbox"/> Oryzias latipes strain Hd-rR chromosome 1 sequence	Oryzias latipes	458	15231	9%	5e-123	77.24%	37713152	CP020665.1
<input checked="" type="checkbox"/> Oryzias latipes strain Hd-rR chromosome 13 sequence	Oryzias latipes	453	21196	9%	6e-122	76.02%	33825776	CP020677.1
<input checked="" type="checkbox"/> Takifugu rubripes genome assembly chromosome: 22	Takifugu rubripes	451	4631	9%	7e-121	76.89%	16056980	LR584243.1
<input checked="" type="checkbox"/> Oryzias latipes strain HNI chromosome 17	Oryzias latipes	445	13849	9%	3e-119	75.88%	28809336	CP020795.1
<input checked="" type="checkbox"/> Oryzias latipes strain HNI chromosome 1	Oryzias latipes	444	15489	9%	1e-118	77.19%	34611496	CP020779.1
<input checked="" type="checkbox"/> Oryzias latipes strain Hd-rR chromosome 6 sequence	Oryzias latipes	444	5006	9%	1e-118	76.15%	32246747	CP020670.1

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ  PRODUCTON
BLAST Sequence Analysis	Effective Date: 02/15//2021 Checked by A. Siclair	Revision #1.0 M. Guzie



6.1.8. The 3rd option for more dissimilar sequences (discontiguous megablast) can be used to find genes with even more variation from the original sequence; yields results without contigs (see DEFINITIONS).

6.2. Protein BLAST

- 6.2.1 Begin with another primary test by running a BLAST for the human myoglobin protein. The amino acid sequence of human myoglobin can be found here: <https://www.ncbi.nlm.nih.gov/protein/P02144.2> (Scroll to the bottom where it says ORIGIN)
- 6.2.2 Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi> to perform the BLAST search for human myoglobin. Select Protein BLAST.
- 6.2.3 Paste the amino acid sequence into the “Query Sequence” box.
- 6.2.4 Make sure the database is set to “Non-redundant protein sequences (nr)”.
- 6.2.5 Under program selection, there are even more algorithms and different ways to analyze protein sequences than for the nucleotide blast. Use the original protein BLAST, blastp, to analyze protein sequence similarities at the most basic level. As with the nucleotide BLAST, proteins with similar sequences will appear with statistical values to indicate their relevance.
- 6.2.6 Return to the BLAST homepage to run an additional blastp on myoglobin. This time, you will pretend you do not know what protein you are searching for or that you only have a segment of the sequence and use blastp to help you determine what type of protein you have. To do this, change the database from non-redundant protein sequences (nr) to an annotated protein database, such as RefSeq Select Proteins or UniProtKB/SwissProt. These databases include annotations which include functions of certain recurring sequences or domains.
- 6.2.7 Delete a portion of the myoglobin amino acid sequence – leaving about 1/3 of a consecutive sequence remaining, and then BLAST it using either RefSeq or UniProtKB/SwissProt as the database.
- 6.2.8 Observe what matches come up to help identify the protein based on only a portion of the sequence.

7. TROUBLE SHOOTING



- 7.1. “No sequence similarity found.” This may be due to a sequence being too short and the program finding too high of an E value. The E-value threshold can be changed under “Algorithm parameters.” Change it to >10 as it is normally capped at 10.
- 7.2. The E value (the expect value) can be thought of as the number of hits one can expect to occur due to chance. The lower the value, the less likely the similar

	<p style="text-align: center;">SKILLS CENTER STANDARD OPERATING PROCEDURE</p>	<p style="text-align: center;">A BIOFIZZ</p>  <p style="text-align: center;">PRODUCTON</p>
<p style="text-align: center;">BLAST Sequence Analysis</p>	<p style="text-align: center;">Effective Date: 02/15//2021 Checked by A. Siclair</p>	<p style="text-align: center;">Revision #1.0 M. Guzie</p>

sequences identified were found just due to chance; they are more likely to be significant.

8. REFERENCES

- BLAST (Basic Local Alignment Search Tool.) National Center for Biotechnology Information. Retrieved from <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
- Heather, J., & Chain, B. (2016, January). The sequence of sequencers: The history of sequencing DNA. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/>
- Lobo, I. (2008) Basic Local Alignment Search Tool (BLAST). *Nature Education* 1(1):215. Retrieved from <https://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096/>.
- Nucleotide Sequence Search. National Center for Biotechnology Information. Retrieved from <https://www.ncbi.nlm.nih.gov/nucleotide/>.
- Protein Sequence Search. National Center for Biotechnology Information. Retrieved from <https://www.ncbi.nlm.nih.gov/protein/>.

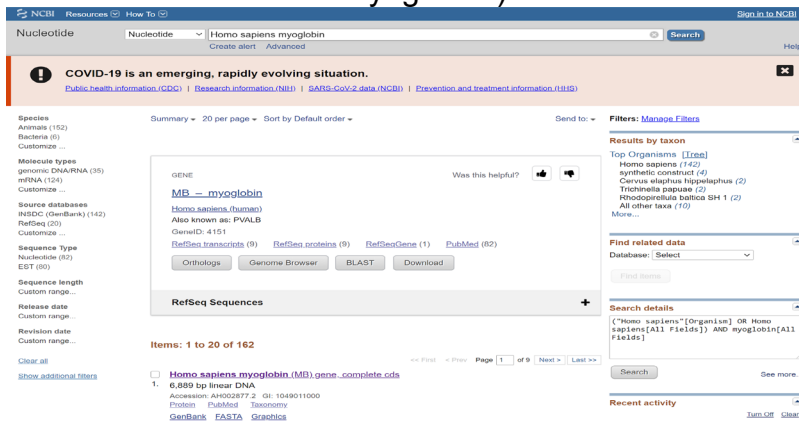
	<p style="text-align: center;">SKILLS CENTER STANDARD OPERATING PROCEDURE</p>	<p style="text-align: center;">A BIOFIZZ</p>  <p style="text-align: center;">PRODUCTON</p>
<p style="text-align: center;">BLAST Sequence Analysis</p>	<p style="text-align: center;">Effective Date: 02/15//2021 Checked by A. Siclair</p>	<p style="text-align: center;">Revision #1.0 M. Guzie</p>

9. MODULE MASTERY TASK

Perform a BLAST sequence analysis search on a gene/protein of your choosing and analyze the results.

9.1. Nucleotide Blast

9.1.1. Choose a gene of interest and find the nucleotide sequence of the gene. Use <https://www.ncbi.nlm.nih.gov/nucleotide/> to find the nucleotide sequence of a gene of interest. Several sequence results may show up even if you specified the organism; make sure you choose the complete gene for linear DNA (Example below of what was selected to find the DNA sequence of human myoglobin.)



The screenshot shows the NCBI Nucleotide search page. The search term is 'Homo sapiens myoglobin'. The results page displays the RefSeq entry for 'MB - myoglobin' (Gene ID: 4151). The entry includes details such as 'Homo sapiens (human)', 'Also known as: PVALB', and 'GeneID: 4151'. The 'RefSeq Sequences' section shows one item: 'Homo sapiens myoglobin (MB) gene, complete cds', which is 6,889 bp linear DNA with accession number AF002877.2. The page also includes navigation options like 'Orthologs', 'Genome Browser', 'BLAST', and 'Download'.



9.1.2. Run both a megablast and a blastn for a gene of choice and make record of at least one gene found from each result and what the values indicated about the similarity between the query gene and the match gene. For the blastn, indicate a homolog (see DEFINITIONS) for the gene from a different species.

9.2. Protein BLAST

9.2.1. Choose a protein of interest and find the amino acid sequence. Use <https://www.ncbi.nlm.nih.gov/protein/> to find the amino acid sequence of a protein of interest. It may be most beneficial to use the protein encoded by the same gene you chose for the nucleotide BLAST to observe characteristics of gene expression and homology.

9.2.2. Run a blastp with the non-redundant protein sequences (nr) database and make record of a several proteins with high sequence similarity.

9.2.3. Delete a portion of the protein and run a blastp using RefSeq or UnitProtKB/SwissProt as the database and make a record of what

	<p style="text-align: center;">SKILLS CENTER STANDARD OPERATING PROCEDURE</p>	<p style="text-align: center;">A BIOFIZZ</p>  <p style="text-align: center;">PRODUCTON</p>
<p style="text-align: center;">BLAST Sequence Analysis</p>	<p style="text-align: center;">Effective Date: 02/15//2021 Checked by A. Siclair</p>	<p style="text-align: center;">Revision #1.0 M. Guzie</p>

domains/polypeptide sequences showed up to help you identify your protein. Also make a note of any polypeptides that may have shown up that indicate a protein of similar function to your protein, or homologs from another organism.