

# TEACHING THE GENOME GENERATION

*BLAST Tutorials*



## BLAST Tutorials

Use the following menu to link to a specific topic within this tutorial series:

- A. [Identifying a gene](#)
- B. [Comparing two or more DNA sequences using BLASTn](#)
- C. [Comparing two or more protein sequences using BLASTp](#)
- D. [Tutorial Sequences](#)

## A. Identifying a gene

This section of the tutorial explains how to take an unknown sequence and identify from which organism and where in the genome it originates.

1. Navigate to [BLAST](#) hosted by the National Center for Biotechnology Information (NCBI).
2. Select “Nucleotide BLAST” under Web BLAST.
3. There is a large box under “Enter Query Sequence.” Copy and paste the **Unknown DNA Sequence** from the [Tutorial Sequences](#) below into the “Enter Query Sequence” box.

BLASTN programs search nucleotide

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file  No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

When entering sequences into BLAST, copy and paste all the text including the description line, which starts with a “>” sign.

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file  No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

```
>Unknown gene
atggagccgcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggctc
gggtagaggaggtcgggcgctgctggagggcgggcgctgcccacgccgaatagttacggtcggag
gccgatccaggtcatgatgatggcagcggccggagtgccggagctgctgctccacggcggagccca
```

4. Leave all the settings as is and scroll down and hit the blue BLAST button.

**BLAST** Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

**+ Algorithm parameters**

5. The page will refresh several times while the alignment job runs. It should take about 30 seconds to load the result.
6. Once the results load, you will see a box labeled “Sequences producing significant alignments” listing genes and organisms that have sequences that match this one.
7. For this alignment, you should notice that the top hit is the “Homo sapiens cyclin dependent kinase inhibitor 2A (CDKN2A),” which is a human gene that is frequently mutated in melanoma skin cancer among other cancers. There is a 100% match to this gene, as indicated by the entry in the “percent identity” column.

Sequences producing significant alignments		Download	New Select columns			
<input checked="" type="checkbox"/> select all 100 sequences selected		GenBank	Graphics	Distance tree of re		
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident
<input checked="" type="checkbox"/> <a href="#">Homo sapiens cyclin dependent kinase inhibitor 2A (CDKN2A), transcript variant 1, mRNA</a>	<a href="#">Homo sapiens</a>	870	870	100%	0.0	100.00%

## B. Comparing more than two DNA sequences using BLASTn

This section of the tutorial will explain how to compare and identify variation between two or more DNA sequences. See also, [the BLAST Tutorial Series: Comparing two or more DNA sequences](#) video.

1. Navigate to [BLAST](#) hosted by the National Center for Biotechnology Information (NCBI).
2. Select “Nucleotide BLAST” under Web BLAST.
3. Check the box “Align two or more sequences” to load a second query box “Enter Subject Sequence.”

The screenshot shows the 'Enter Query Sequence' form on the NCBI BLAST website. The form includes a text input for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Query subrange' section with 'From' and 'To' fields, an 'Or, upload file' section with a 'Choose File' button, and a 'Job Title' field. At the bottom, the checkbox 'Align two or more sequences' is checked, and a purple arrow points to it.

To identify variation in different sequences, those sequences must be compared to a standard sequence called a *reference sequence*. This standard sequence is a point of reference for a specific gene and will indicate if variation in a gene sequence has occurred.

4. There is a large box under “Enter Query Sequence.” Copy and paste the [DNA Reference Sequence](#) from the [Tutorial Sequences](#) below into the “Enter Query Sequence” box. Make sure to copy and paste the entire sequence including the description line starting with “>” that comes directly before the DNA sequence.

The screenshot shows the 'Enter Query Sequence' form with the following content:

- Enter accession number(s), gi(s), or FASTA sequence(s):** >DNA Reference Sequence  
atggagcggcggggagcagcatggagcctcggctgactggctggccacggcgcggccgggggtc  
gggtagaggaggtgcggcgctgctggagcggggcgctgcccacgcaccgaatagtlacggtcggag  
gccgatccaggatcatgatgggcagcggcggagctgctgctccacggcggagccca
- Job Title:** DNA Reference Sequence
- Align two or more sequences:**

- Then copy and paste the sequences **DNA Sequence 1**, **DNA Sequence 2**, and **DNA Sequence 3** from the [Tutorial Sequences](#) found under “Other DNA Sequences” below into the box labeled “Enter Subject Sequence.” You can simply copy and paste all three sequences at once starting with the > sign preceding the first sequence. Leaving the sequence description lines will help you distinguish between the different sequences once they have been aligned to the reference sequence.

**Enter Subject Sequence**
Clear Subject subrange ?

Enter accession number(s), gi(s), or FASTA sequence(s) ?

```
>DNA Sequence 1
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcccgggggc
gggtagaggagggtgcgggcgctgctggagcggggcgctgcccaacgcaccgaatgttacggtcggag
gccgatccagggtcatgatgagggcagcgcggcggagctgctgctccacggcggagccca
actgcccgaccccaccactctcaccgacccgacgacgctgcccgggagggcttcctggacacgctgg
tggctgcaccgggcccggggcgcgctggacgtgcgcgatgctggggccgctgcccgtggactggctg
aggagctggccatcgcgatgtcgcacggtacctgcgcggctgcccggggcaccagaggcagtaacct
gccgcatalagatgccggaaggctccctcagacatcccgaattga

>DNA Sequence 2
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcccgggggc
gggtagaggagggtgcgggcgctgctggagcggggcgctgcccaacgcaccgaatgttacggtcggag
gggtagaggagggtgcgggcgctgctggagcggggcgctgcccaacgcaccgaatgttacggtcggag
```

Or, upload file
Choose File
No file chosen
?

- Keeping all the other settings as is, use the BLAST button to compare the sequences and wait for the page to refresh with results.

BLAST

Search database **Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**  
 Show results in a new window

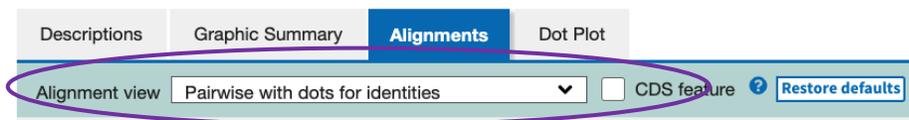
+ Algorithm parameters

- Once the results load, the “Descriptions” tab should include three alignments, one for **DNA Sequence 1**, one for **DNA Sequence 2**, and one for **DNA Sequence 3** each compared to the **DNA Reference Sequence**.

Descriptions		Graphic Summary	Alignments							
Sequences producing significant alignments				Download	Select columns	Show	100			
<input checked="" type="checkbox"/> select all <span style="font-weight: normal;">3 sequences selected</span>				<a href="#">Graphics</a> <a href="#">Distance tree of results</a> <a href="#">MSA Viewer</a>						
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	
<input checked="" type="checkbox"/>	DNA Sequence 1		870	870	100%	0.0	100.00%	471	Query_62939	
<input checked="" type="checkbox"/>	DNA Sequence 3		865	865	100%	0.0	99.79%	471	Query_62941	
<input checked="" type="checkbox"/>	DNA Sequence 2		841	841	100%	0.0	98.74%	476	Query_62940	

The column labeled “Query Cover” provides an indication of the length of each subject sequence compared to the reference sequence. 100% coverage indicates that the subject sequence spans the entire length of the **DNA Reference Sequence**. The column labeled “Per. Ident” stands for *percent identity*, which is the percentage of the nucleotides that are the same between the two sequences. 100% indicates that at each position of the alignment, the nucleotide in the subject sequence is identical to the reference sequence. A percent identity below 100% indicates that there are differences between the two sequences meaning there could be base exchanges (a different base a position), deletions, or insertions within the subject sequence.

8. To see the comparison of the sequences themselves, click on the “Alignments” tab.
9. The easiest way to visualize the alignment of the sequences is to select “Pairwise with dots for identities” as the “Alignment view.”



In this view, each subject sequence is aligned to the **Reference Sequence** individually. The specific subject for each comparison is indicated at the top of the alignment. The numbers at the start and end of each line represent the nucleotide number in the sequence starting with 1 on the first line.

[Download](#) [Graphics](#)

**DNA Sequence 1**  
Sequence ID: **Query\_62939** Length: **471** Number of Matches: **1**

Range 1: 1 to 471 [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
870 bits(471)	0.0	471/471(100%)	0/471(0%)	Plus/Plus

```

Query 1  ATGGAGCCGGCGCGGGGAGCAGCATGGAGCCTTCGGCTGACTGGCTGGCCACGGCCGCG 60
Sbjct 1  .....

```

Any bases that are the same between the subject sequence (either **DNA Sequence 1, 2, or 3**) and the query sequence (**DNA Reference Sequence**) are represented by a dot. Those that differ are listed as a red letter or a dash.

In this example, **DNA Sequence 1** does not have any differences from the **DNA Reference Sequence**, but **DNA Sequences 2 and 3** have differences.

In the alignment with **DNA Sequence 3**, there is a red T in position 353 in the sequence. This indicates that there is a single base substitution from a C in the **Reference Sequence** to a T in **DNA Sequence 3**.

```

Query 301  GGGGCGCGGCTGGACGTGCGCGATGCCTGGGGCCGCTGCCCCGTTGGACCTGGCTGAGGAG 360
Sbjct 301  .....T..... 360

```

In the alignment with [DNA Sequence 2](#), there is also a red T in position 353 in the sequence. This indicates that there is a single base substitution from a C in the [DNA Reference Sequence](#) to a T in [DNA Sequence 2](#). There is also a group of red Ts in the subject sequence and dashes in the query sequence. This indicates that there is an insertion of DNA bases in [DNA Sequence 2](#) that are not present in the [DNA Reference Sequence](#) and the dashes are place holders for the missing nucleotides. Had the dashes been in the subject sequence, this would indicate a deletion in that sequence and the dashes represent missing nucleotides that are present in the [DNA Reference Sequence](#).

```

Query  181  GAGCTGCTGCTGCTCCACGGCGCGGAGCCCAAC-----TGCGCCGACCCCGCCACTCTCA  235
Sbjct  181  .....TTTTT.....  240

Query  236  CCCGACCCGTGCACGACGCTGCCCGGGAGGGCTTCCTGGACACGCTGGTGGTGTGCACC  295
Sbjct  241  .....  300

Query  296  GGGCCGGGGCGCGGCTGGACGTGCGCGATGCCTGGGGCCGTCTGCCCGTGGACCTGGCTG  355
Sbjct  301  .....T...  360

```

### C. Comparing more than two protein sequences using BLASTp

This tutorial will explain how to compare and identify variation between two or more protein sequences. See also, the [BLAST Tutorial Series: Comparing two or more protein sequences](#) video.

1. Navigate to [BLAST](#) hosted by the National Center for Biotechnology Information (NCBI).
2. Select “Protein BLAST” under Web BLAST.
3. Check the box “Align two or more sequences” to load a second query box “Enter Subject Sequence.”

To identify variation in different sequences, those sequences must be compared to a standard sequence called a *reference sequence*. This standard sequence is a point of reference for a specific gene and will indicate if variation in a protein sequence has occurred.

4. There is a large box under “Enter Query Sequence.” Copy and paste the **Protein Reference Sequence** from the [Tutorial Sequences](#) below into the “Enter Query Sequence” box. Make sure to copy and paste the entire sequence including the description line starting with “>” that comes directly before the DNA sequence.

5. Then copy and paste the sequences **Protein Sequence 1** and **Protein Sequence 2** from the [Tutorial Sequences](#) found under “Other Protein Sequences” below into the box labeled “Enter Subject Sequence.” You can simply copy and paste all three sequences at once starting with the > sign preceding the first sequence. Leaving the sequence description lines will help you distinguish between the two sequences once they have been aligned to the reference sequence.

6. Keeping all the other settings as is, use the BLAST button to compare the sequences and wait for the page to refresh with results.

7. Once the results load, the “Descriptions” tab should include two alignments, one for **Protein Sequence 1** and one for **Protein Sequence 2** compared to the **Protein Reference Sequence**.

Descriptions Graphic Summary Alignments

Sequences producing significant alignments Download Select columns Show 100

select all 2 sequences selected 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
<input checked="" type="checkbox"/> Protein Sequence 1		954	954	100%	0.0	99.79%	472	Query_1370
<input checked="" type="checkbox"/> Protein Sequence 2		950	950	100%	0.0	99.79%	471	Query_1371

The column labeled “Query Cover” provides an indication of the length of each subject sequence compared to the reference sequence. 100% coverage indicates that the subject sequence spans the entire length of the **Reference Sequence**. The column labeled “Per. Ident” stands for percent identity, which is the percentage of the amino acids that are the same between the two sequences. 100% indicates that at each position of the alignment, the amino acid in the subject sequence is identical to the reference sequence. A percent identity below 100% indicates that there are differences between the two sequences meaning there could be amino acid substitutions (a different amino acid in one position) as well as deletions or insertions of amino acids within the subject sequence.

- To see the comparison of the sequences themselves, click on the “Alignments” tab.
- The easiest way to visualize the alignment of the sequences is to select “Pairwise with dots for identities” as the “Alignment view.”

Descriptions Graphic Summary **Alignments**

Alignment view Pairwise with dots for identities Restore defaults

In this view, each subject sequence is aligned to the **Protein Reference Sequence** individually. The specific subject for each comparison is indicated at the top of the alignment. The numbers at the start and end of each line represent the amino acid number in the sequence starting with 1 on the first line.

Download Graphics

**Protein Sequence 1**  
Sequence ID: Query\_1370 Length: 472 Number of Matches: 1

Range 1: 1 to 472 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
954 bits(2465)	0.0	Compositional matrix adjust.	471/472(99%)	471/472(99%)	0/472(0%)

```

Query 1  MLGTVKMEGHETS D WNSY YADTQ EAYSSVPVSNMNSGLGSMNSMNTYMTMNTMTTSGNMT 60
Sbjct 1  .....

```

Any amino acids that are the same between the subject sequence (either **Protein Sequence 1** or **2**) and the query sequence (**Protein Reference Sequence**) are represented by a dot. Those that differ are listed as a red letter or a dash.

In this example, both **Protein Sequences 2** and **3** have differences from the reference.

In the alignment with **Protein Sequence 1**, there is a red R in position 247 in the sequence. This indicates that there is a single amino acid substitution from a H (histidine) in the **Protein Reference Sequence** to an R (arginine) in **Protein Sequence 1**.

```
Query 241 GSYWTLHPDSGNMFENGCYLRRQKRFKCEKQPGAGGGGGSGSGGSGAKGGPESRKDPSGA 300
Sbjct 241 .....R..... 300
```

In the alignment with **Protein Sequence 2**, there is a dash in the subject sequence. This indicates that there is a deletion in **Protein Sequence 2** for an amino acid that is present in the **Protein Reference Sequence**. A dash is a placeholder for missing amino acids. Had the dash been in the query sequence, this would indicate an insertion in the subject protein sequence and the dash represents an additional amino acid that is not present in the **Protein Reference Sequence**.

```
Query 241 GSYWTLHPDSGNMFENGCYLRRQKRFKCEKQPGAGGGGGSGSGGSGAKGGPESRKDPSGA 300
Sbjct 241 .....-..... 299
```

## D. Tutorial Sequences

### Unknown Gene Sequence

>Unknown gene

```
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcgggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggatcatgatgatgggcagcgcgccgagtgccggagctgctgctgctccacggcgcgaggccc
aactgcgccgacccccgcaactctcaccgcaccgctgcacgacgctgcccgggagggcttcctggacacgc
tgggtggtgctgcaccgggcccggggcgcggtggacgtgcgcgatgcctggggccgctctgcccgctggacct
ggctgaggagctgggcatcgcgatgtcgcacgggtacctgcgcgcggctgcggggggcaccagaggcagt
aaccaatgcccgcataagatgccgcggaaggtccctcagacatccccgattga
```

### DNA Reference Sequence

>DNA Reference Sequence

```
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcgggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggatcatgatgatgggcagcgcgccgagtgccggagctgctgctgctccacggcgcgaggccc
aactgcgccgacccccgcaactctcaccgcaccgctgcacgacgctgcccgggagggcttcctggacacgc
tgggtggtgctgcaccgggcccggggcgcggtggacgtgcgcgatgcctggggccgctctgcccgctggacct
ggctgaggagctgggcatcgcgatgtcgcacgggtacctgcgcgcggctgcggggggcaccagaggcagt
aaccaatgcccgcataagatgccgcggaaggtccctcagacatccccgattga
```

### Other DNA Sequences

>DNA Sequence 1

```
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcgggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggatcatgatgatgggcagcgcgccgagtgccggagctgctgctgctccacggcgcgaggccc
aactgcgccgacccccgcaactctcaccgcaccgctgcacgacgctgcccgggagggcttcctggacacgc
tgggtggtgctgcaccgggcccggggcgcggtggacgtgcgcgatgcctggggccgctctgcccgctggacct
ggctgaggagctgggcatcgcgatgtcgcacgggtacctgcgcgcggctgcggggggcaccagaggcagt
aaccaatgcccgcataagatgccgcggaaggtccctcagacatccccgattga
```

>DNA Sequence 2

```
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcgggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggatcatgatgatgggcagcgcgccgagtgccggagctgctgctgctccacggcgcgaggccc
aacttttttgcgccgacccccgcaactctcaccgcaccgctgcacgacgctgcccgggagggcttcctgga
cacgctggtggtgctgcaccgggcccggggcgcggtggacgtgcgcgatgcctggggccgctctgcccgctg
gacctggttagaggagctgggcatcgcgatgtcgcacgggtacctgcgcgcggctgcggggggcaccagag
gcagtaaccaatgcccgcataagatgccgcggaaggtccctcagacatccccgattga
```

>DNA Sequence 3

```

Atggagccggcgggcggggagcagcatggagccttcggctgactggctggccacggccgcgggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggatcatgatgatgggcagcgcccgagtgccggagctgctgctgctccacggcgcgagccc
aactgcgccgacccccgactctcacccgacccgtgcacgacgctgcccgggagggcttcctggacacgc
tggtggtgctgcaccgggcccggggcgcggtggacgtgcgcgatgctggggccgtctgcccggtggacct
ggttgaggagctgggcatcgcgatgctgcacgggtacctgcgcgcggtgccccggggcaccagaggcagt
aaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga

```

## Protein Reference Sequence

>Protein Reference Sequence

```

Mlgtvkmeghetsdwnsyadtqeaayssvpvsnmns glsmnsmntytmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgm gamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsragggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtlhpds gnmfengcylrrqkrfkcekqpgagggggs
gsggsgakggpesrkdpsgasnpsadsplhrghv hktgqlegapapppaaspqtl dhsgatatggaselk
tpasstappissgpgalasvpashpahglaphesqlh lkgdphysfnhpfsinnlmsseqqhkldfkay
eqalqyspygstl paslplgsasvttrspiepsalepayyqgvysr pvlnts

```

## Other Protein Sequences

>Protein Sequence 1

```

mlgtvkmeghetsdwnsyadtqeaayssvpvsnmns glsmnsmntytmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgm gamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsragggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtl rpdsgnmfengcylrrqkrfkcekqpgagggggs
gsggsgakggpesrkdpsgasnpsadsplhrghv hktgqlegapapppaaspqtl dhsgatatggaselk
tpasstappissgpgalasvpashpahglaphesqlh lkgdphysfnhpfsinnlmsseqqhkldfkay
eqalqyspygstl paslplgsasvttrspiepsalepayyqgvysr pvlnts

```

>Protein Sequence 2

```

mlgtvkmeghetsdwnsyadtqeaayssvpvsnmns glsmnsmntytmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgm gamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsragggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtl hds gnmfengcylrrqkrfkcekqpgaggggsg
sggsgakggpesrkdpsgasnpsadsplhrghv hktgqlegapapppaaspqtl dhsgatatggaselk
passtappissgpgalasvpashpahglaphesqlh lkgdphysfnhpfsinnlmsseqqhkldfkaye
qalqyspygstl paslplgsasvttrspiepsalepayyqgvysr pvlnts

```