TEACHING THE GENOME GENERATION

BLAST Tutorials

The Jackson Laboratory
Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

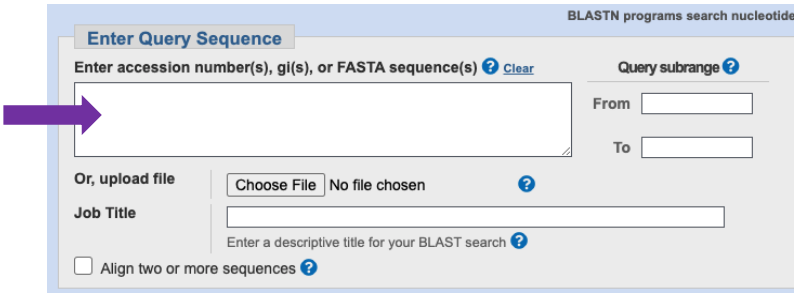
BLAST Tutorials

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

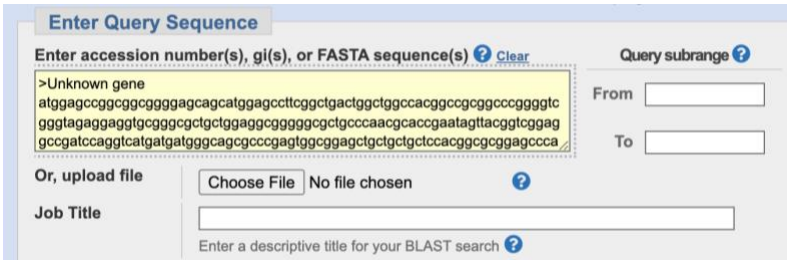
A. Identifying a gene

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.

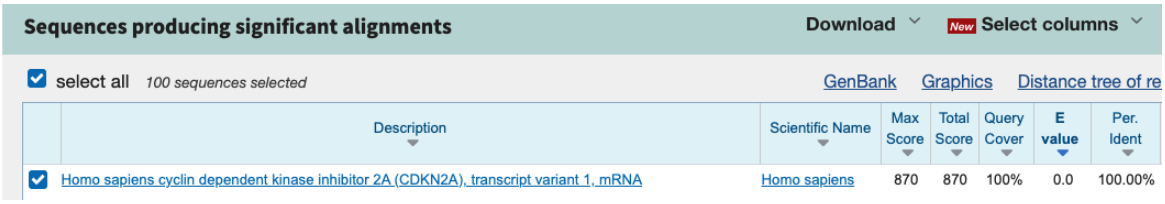


When entering sequences into BLAST, copy and paste all the text including the description line, which starts with a ">" sign.



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

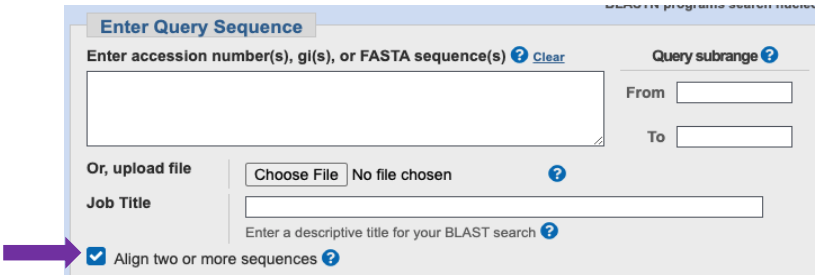
5. The page will refresh several times while the alignment job runs. It should take about 30 seconds to load the result.
6. Once the results load, you will see a box labeled "Sequences producing significant alignments" listing genes and organisms that have sequences that match this one.
7. For this alignment, 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.



B. Comparing more than two DNA sequences using BLASTn

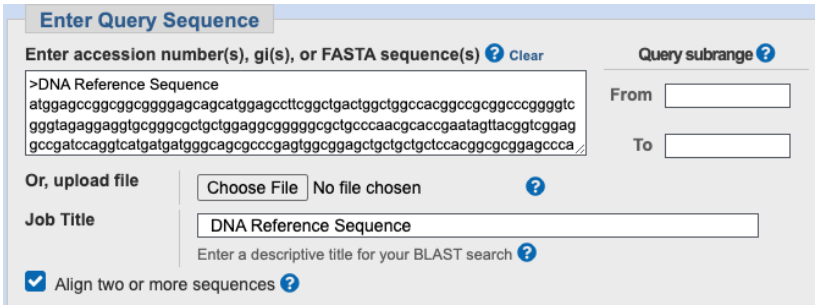
This section of the tutorial will explain how to compare and identify variation between two or more DNA sequences. See also, the BLAST Tutorial Series: Comparing two or more DNA sequences video.

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."

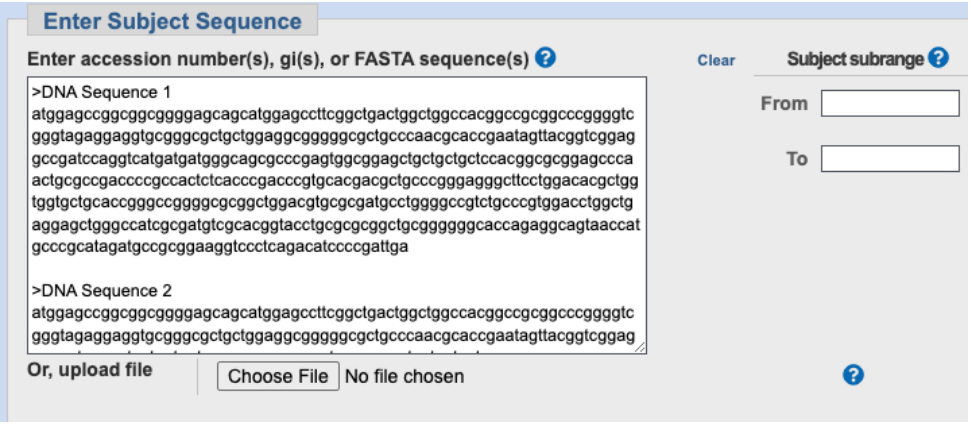


To identify variation in different sequences, those sequences must be compared to a standard sequence called a *reference sequence*. This standard sequence is a point of reference for a specific gene and will indicate if variation in a gene sequence has occurred.

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.

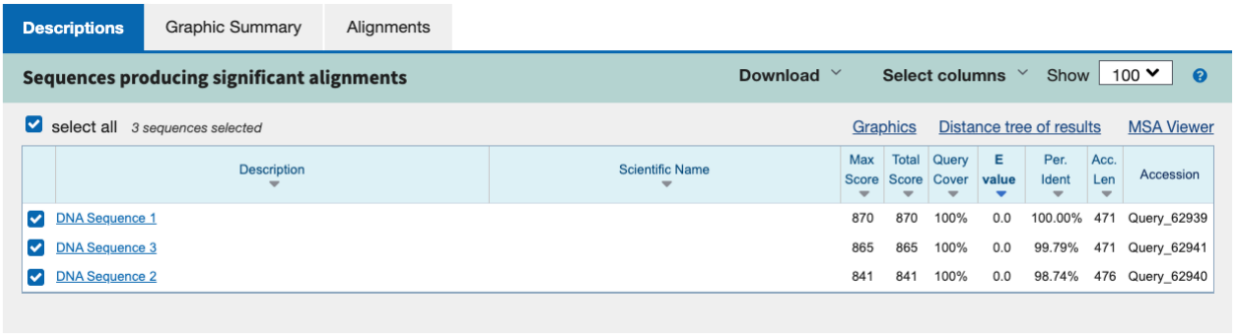


5. Then copy and paste the sequences **DNA Sequence 1**, **DNA Sequence 2**, and **DNA Sequence 3** from the Tutorial Sequences found under "Other DNA Sequences" below into the box labeled "Enter Subject Sequence." You can simply copy and paste all three sequences at once staring with the > sign proceeding 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.



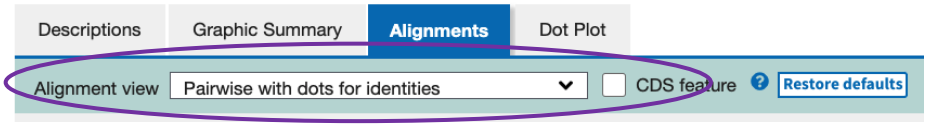
6. Keeping all the other settings as is, use the BLAST button to compare the sequences and wait for the page to refresh with results.

7. Once the results load, the "Descriptions" tab should include 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**.

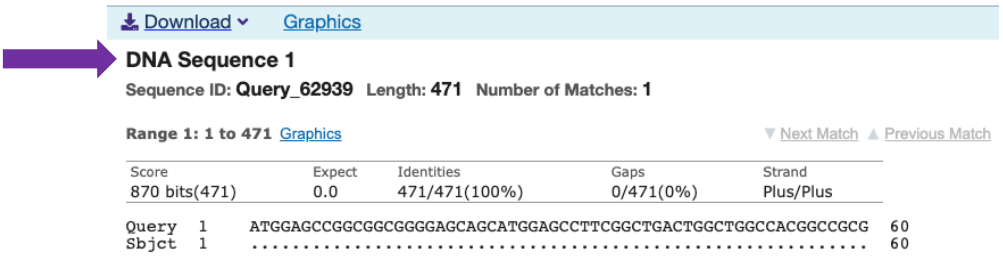


The column labeled “Query Cover” provides an indication of the length of each subject sequence compared to the reference sequence. 100% coverage indicates that the subject sequence spans the entire length of the **DNA Reference Sequence**. The column labeled “Per. Ident” stands for *percent identity*, 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.”



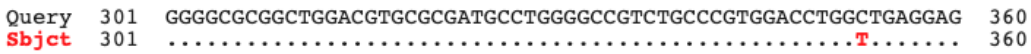
In this view, each subject sequence is aligned to the **Reference Sequence** individually. The specific subject for each comparison is indicated at the top of the alignment. The numbers at the start and end of each line represent the nucleotide number in the sequence starting with 1 on the first line.



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**.



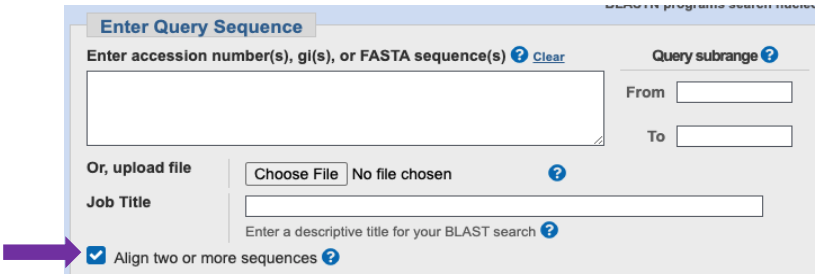
In the alignment with **DNA Sequence 2**, there is also a red T in position 353 in the sequence. This indicates that there is a single base substitution from a C in the **DNA Reference Sequence** to a T in **DNA Sequence 2**. There is also a group of red Ts in the subject sequence and dashes in the query 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  CCCGACCCGTGCACGACGCTGCCCGGGAGGGCTTCCTGGACACGCTGGTGGTGCTGCACC  295
Sbjct  241  ............................................................  300

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

C. Comparing more than two protein sequences using BLASTp

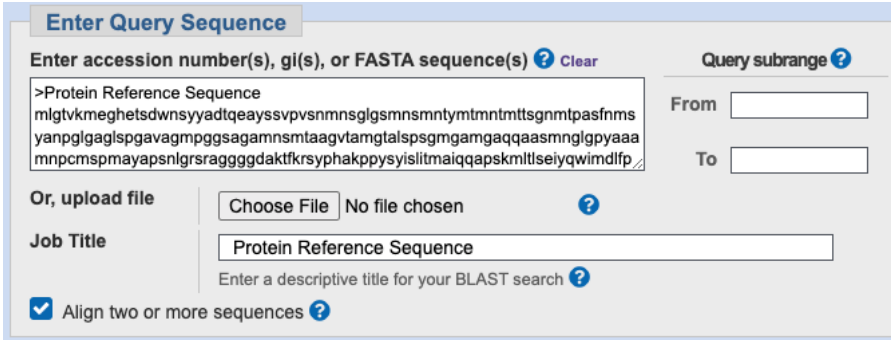
This tutorial will explain how to compare and identify variation between two or more protein sequences. See also, the [BLAST Tutorial Series: Comparing two or more protein sequences] video.

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.”

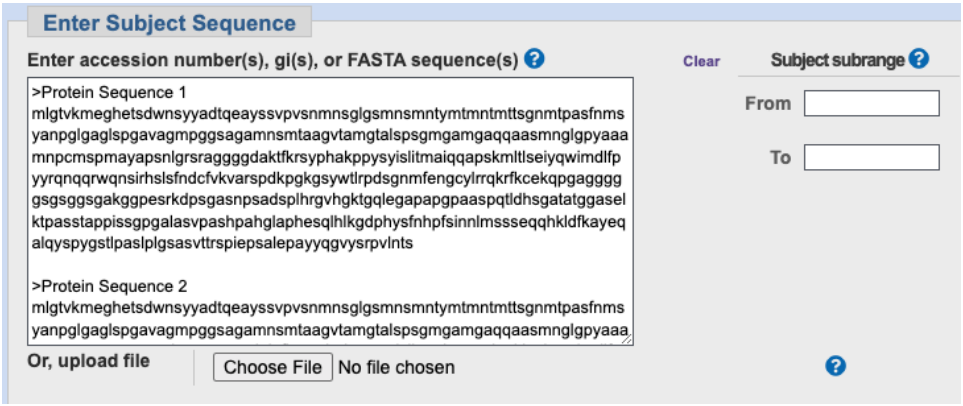


To identify variation in different sequences, those sequences must be compared to a standard sequence called a *reference sequence*. This standard sequence is a point of reference for a specific gene and will indicate if variation in a protein sequence has occurred.

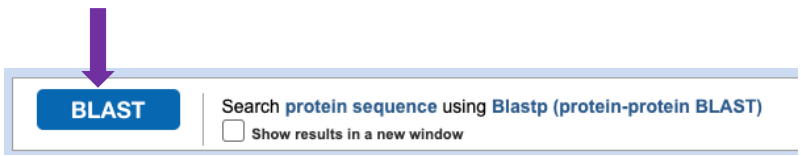
4. There is a large box under “Enter Query Sequence.” Copy and paste the **Protein Reference Sequence** from the Tutorial Sequences below into the “Enter Query Sequence” box. Make sure to copy and paste the entire sequence including the description line starting with “>” that comes directly before the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



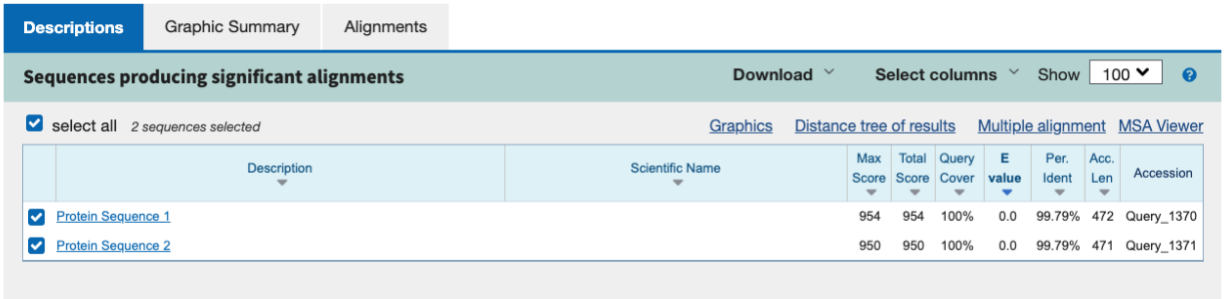
5. Then copy and paste the sequences **Protein Sequence 1** and **Protein Sequence 2** from the Tutorial Sequences found under “Other Protein Sequences” below into the box labeled “Enter Subject Sequence.” You can simply copy and paste all three sequences at once staring with the > sign proceeding the 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.

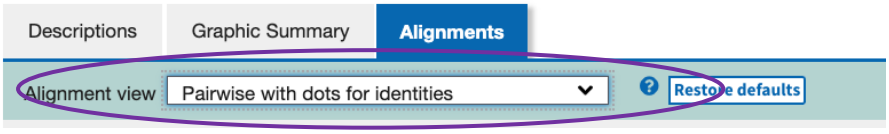


7. Once the results load, the “Descriptions” tab should include two alignments, one for **Protein Sequence 1** and one for **Protein Sequence 2** compared to the **Protein Reference Sequence**.

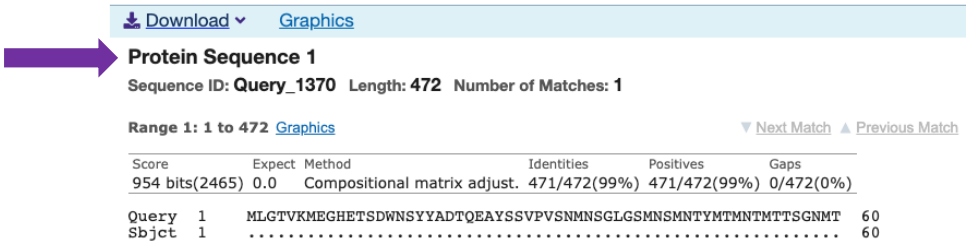


The column labeled "Query Cover" provides an indication of the length of each subject sequence compared to the reference sequence. 100% coverage indicates that the subject sequence spans the entire length of the **Reference Sequence**. The column labeled "Per. Ident" stands for percent identity, which 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.

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 **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.



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 place holder 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
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggtcatgatgatgggcagcgcccgagtggcggagctgctgctgctccacggcgcggagccc
aactgcgccgaccccgccactctcacccgacccgtgcacgacgctgcccgggagggcttcctggacacgc
tggtggtgctgcaccgggccggggcgcggctggacgtgcgcgatgcctggggccgtctgcccgtggacct
ggctgaggagctgggccatcgcgatgtcgcacggtacctgcgcgcggctgcgggggcaccagaggcagt
aaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga
```

DNA Reference Sequence

```
>DNA Reference Sequence
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggtcatgatgatgggcagcgcccgagtggcggagctgctgctgctccacggcgcggagccc
aactgcgccgaccccgccactctcacccgacccgtgcacgacgctgcccgggagggcttcctggacacgc
tggtggtgctgcaccgggccggggcgcggctggacgtgcgcgatgcctggggccgtctgcccgtggacct
ggctgaggagctgggccatcgcgatgtcgcacggtacctgcgcgcggctgcgggggcaccagaggcagt
aaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga
```

Other DNA Sequences

```
>DNA Sequence 1
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggtcatgatgatgggcagcgcccgagtggcggagctgctgctgctccacggcgcggagccc
aactgcgccgaccccgccactctcacccgacccgtgcacgacgctgcccgggagggcttcctggacacgc
tggtggtgctgcaccgggccggggcgcggctggacgtgcgcgatgcctggggccgtctgcccgtggacct
ggctgaggagctgggccatcgcgatgtcgcacggtacctgcgcgcggctgcgggggcaccagaggcagt
aaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga

>DNA Sequence 2
atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggtcatgatgatgggcagcgcccgagtggcggagctgctgctgctccacggcgcggagccc
aacttttttgcgccgaccccgccactctcacccgacccgtgcacgacgctgcccgggagggcttcctggga
cacgctggtggtgctgcaccgggccggggcgcggctggacgtgcgcgatgcctggggccgtctgcccgtg
gacctggttgaggagctgggccatcgcgatgtcgcacggtacctgcgcgcggctgcgggggcaccagag
gcagtaaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga

>DNA Sequence 3
```

```
Atggagccggcggcggggagcagcatggagccttcggctgactggctggccacggccgcggcccggggtc
gggtagaggaggtgcgggcgctgctggaggcgggggcgctgcccaacgcaccgaatagttacggtcggag
gccgatccaggtcatgatgatgggcagcgcccgagtggcggagctgctgctgctccacggcgcggagccc
aactgcgccgaccccgccactctcacccgacccgtgcacgacgctgcccgggagggcttcctggacacgc
tggtggtgctgcaccgggccggggcgcggctggacgtgcgcgatgcctggggccgtctgcccgtggacct
ggttgaggagctgggccatcgcgatgtcgcacggtacctgcgcgcggctgcgggggcaccagaggcagt
aaccatgcccgcatagatgccgcggaaggtccctcagacatccccgattga
```

Protein Reference Sequence

```
>Protein Reference Sequence
Mlgtvkmeghetsdwnsyyadtqeayssvpvsnmnsglgsmnsmntymtmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgmgamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsraggggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtlhpdsgnmfengcylrrqkrfkcekqpgaggggs
gsggsgakggpesrkdpsgasnpsadsplhrgvhgktgqlegapapgpaaspqtldhsgatatggaselk
tpasstappissgpgalasvpashpahglaphesqlhlkgdphysfnhpfsinnlmssseqqhkldfkay
eqalqyspygstlpaslplgsasvttrspiepsalepayyqgvysrpvlnts
```

Other Protein Sequences

```
>Protein Sequence 1
mlgtvkmeghetsdwnsyyadtqeayssvpvsnmnsglgsmnsmntymtmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgmgamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsraggggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtlrpdsgnmfengcylrrqkrfkcekqpgaggggs
gsggsgakggpesrkdpsgasnpsadsplhrgvhgktgqlegapapgpaaspqtldhsgatatggaselk
tpasstappissgpgalasvpashpahglaphesqlhlkgdphysfnhpfsinnlmssseqqhkldfkay
eqalqyspygstlpaslplgsasvttrspiepsalepayyqgvysrpvlnts

>Protein Sequence 2
mlgtvkmeghetsdwnsyyadtqeayssvpvsnmnsglgsmnsmntymtmntmttsgnmtpasfnmsyan
pglgaglspgavagmpggsagamnsmtaagvtamgtalspsgmgamgaqqaasmnglgpyaaamnpcmsp
mayapsnlgrsraggggdaktfkrsyphakppysyislitmaiqqapskmltlseiyqwimdlfpyyrqn
qqrwqnsirhslsfndcfvkvarspdkpgkgsywtlhdsgnmfengcylrrqkrfkcekqpgagggggsg
sggsgakggpesrkdpsgasnpsadsplhrgvhgktgqlegapapgpaaspqtldhsgatatggaselkt
passtappissgpgalasvpashpahglaphesqlhlkgdphysfnhpfsinnlmssseqqhkldfkaye
qalqyspygstlpaslplgsasvttrspiepsalepayyqgvysrpvlnts
```