Non- graded activities

### Seeing a genome sequence and genes

Coat Protein: The gene is located from 1335 to1727

tggct tctaa - First 10 Nucleotides

431 Nucleotides

Replicase(Beta Subunit): The gene is located from 1761 to 3398.

atgtcgaaga - First 10 Nucleotides

1799 Nucleotides

Protein Length:

Coat Protein: MASNFTQFVL - First 10 Amino Acids

142 Amino Acids

Replicase(Beta Subunit): MSKTTKKFNSL - First 10 Amino Acids

731 Amino Acids

### Finding open reading frames

Can you find the start and stop codon in the following sequence?

AATTGCAGTACATGCGATCATTCATCGATGCTAGTAAGTCAGTCGATTAATGCTAGTCAG

Start

Stop

At what positions do the start and stop codons begin, respectively?

Start Codon - 12th position

Stop Codon - 31-33 for 1st TAG, TAA is 48 and TAG is 64

In Genbank format: 60 bp

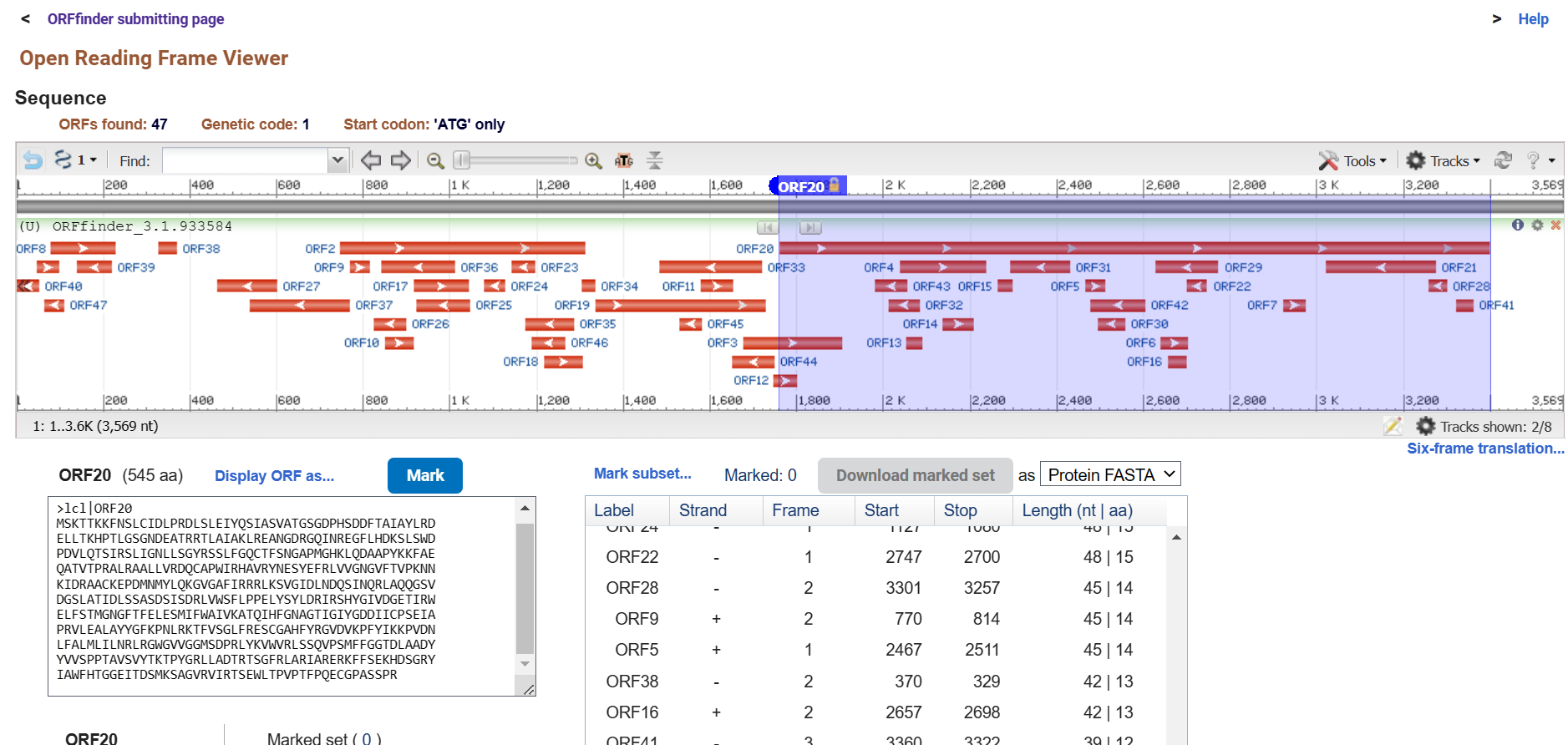
1 AATTGCAGTA CATGCGATCA TTCATCGATG CTAGTAAGTC AGTCGATTAA TGCTAGTCAG

Can you enter the sequence of the complete ORF (with the stop codon)?

ATGCGATCA TTCATCGATG CTAG

Now, imagine doing the same for a much larger genome sequence! A longer genome sequence may have multiple ORFs. It will be better to use a computer program to do this. Interestingly, some bioinformaticians have already written such programs and they are publicly available. Use ORFfinder to detect ORFs in the following sequence - <https://www.ncbi.nlm.nih.gov/nuccore/NC_001417.2?report=fasta>. Try all possible variations of input parameters. Write a detailed note on the results with an explanation of every parameter.

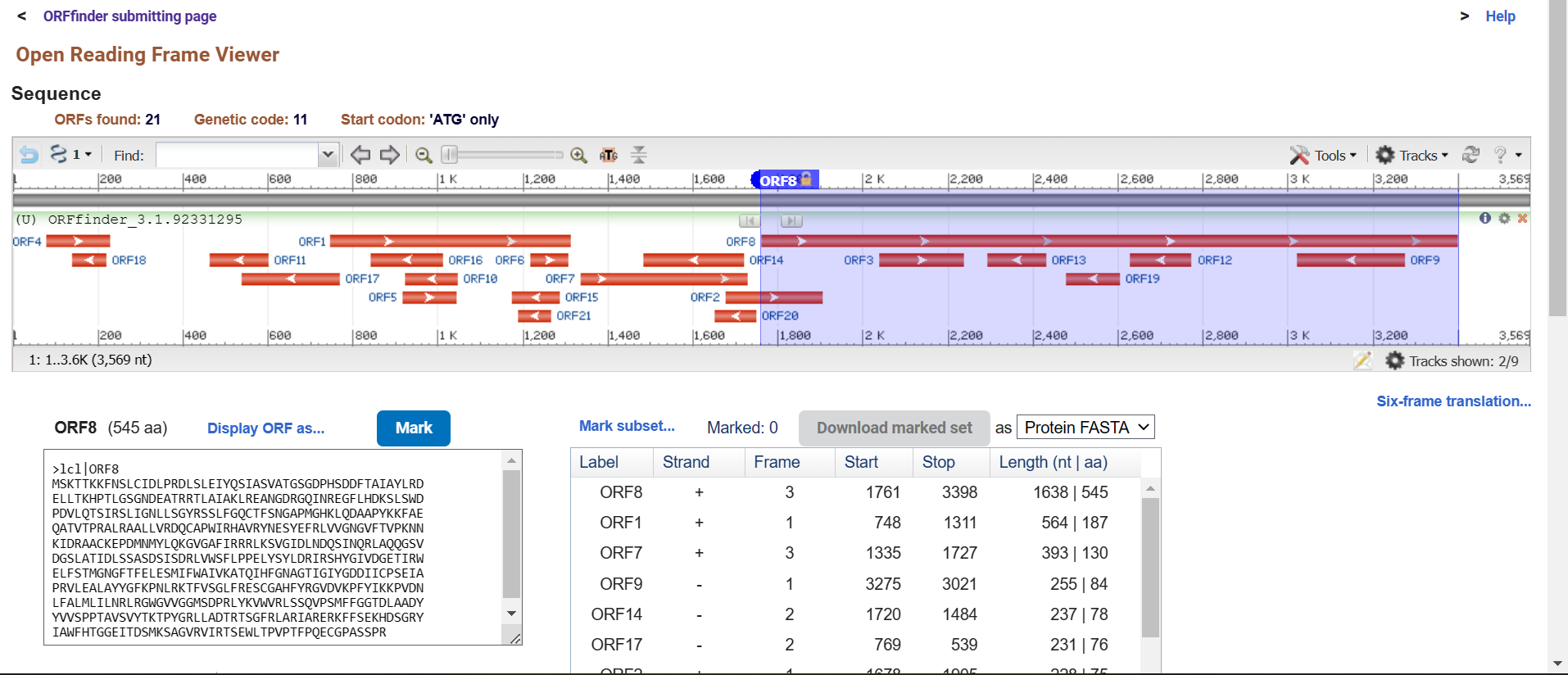
The parameter "Minimal ORF length" sets the minimum nucleotide length required for detecting an ORF. When the organism is unknown, the standard genetic code is used. However, if the organism is known, different organisms may have slightly different translation tables. The genetic code parameter is used to specify the appropriate translation table. For the given phage genome of the prokaryotic virus E. coli, table 11, the Bacterial, Archaeal, and Plant Plastid Code, is used. The "ORF start codon to use" parameter specifies which start codon should be used. If a different genetic code is selected, this parameter can be used to specify whether the corresponding start codons should be used, only ATG, or any sense codon. The "Ignore nested ORFs" parameter specifies whether to ignore nested ORFs, which are ORFs inside another ORF belonging to the same frame.



[ORFfinder Viewer - NCBI (nih.gov)](https://www.ncbi.nlm.nih.gov/orffinder/) - for the given phage genome.

Parameters:

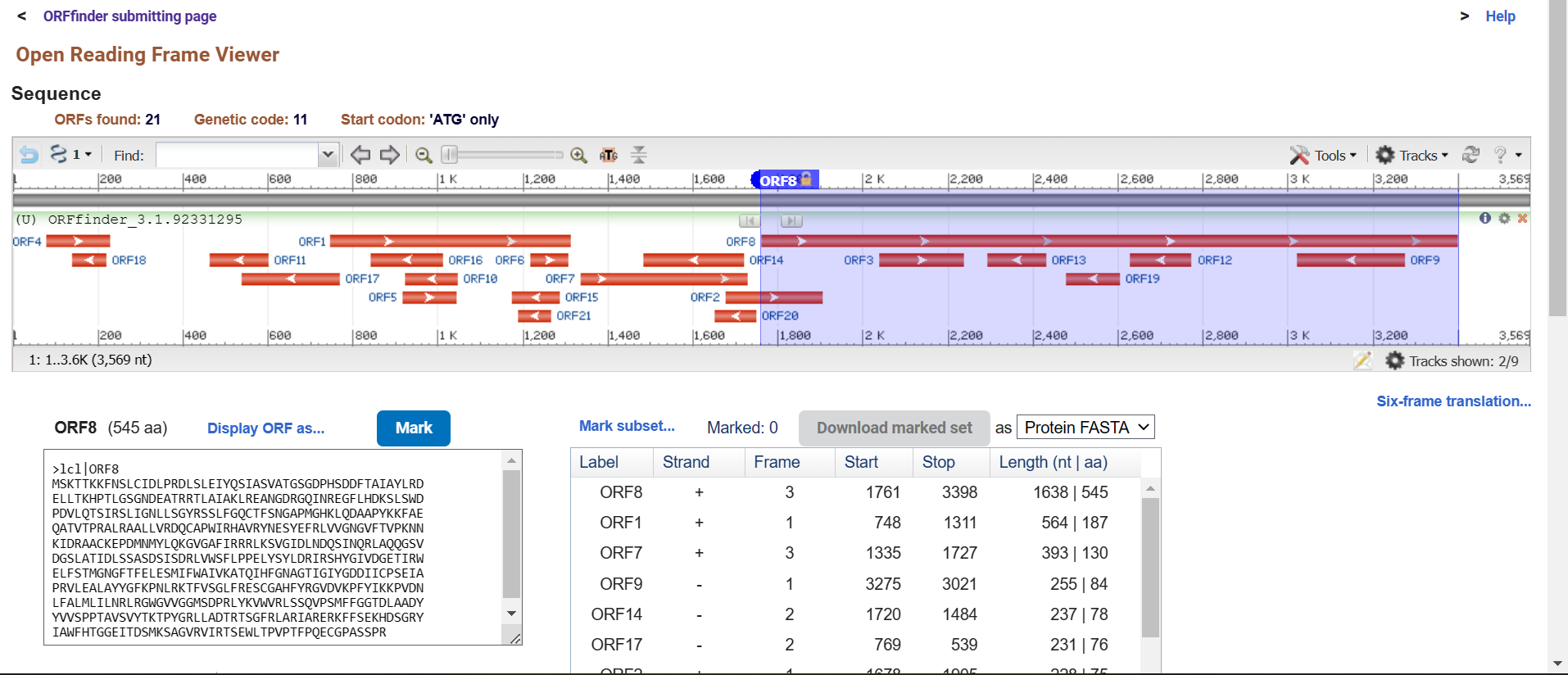
1. Minimal orf length = 75
2. Genetic code = 1 (standard)
3. ORF start codon = ATG only
4. Ignore nested ORFs = no



[ORFfinder Viewer - NCBI (nih.gov)](https://www.ncbi.nlm.nih.gov/orffinder/)

Parameters:

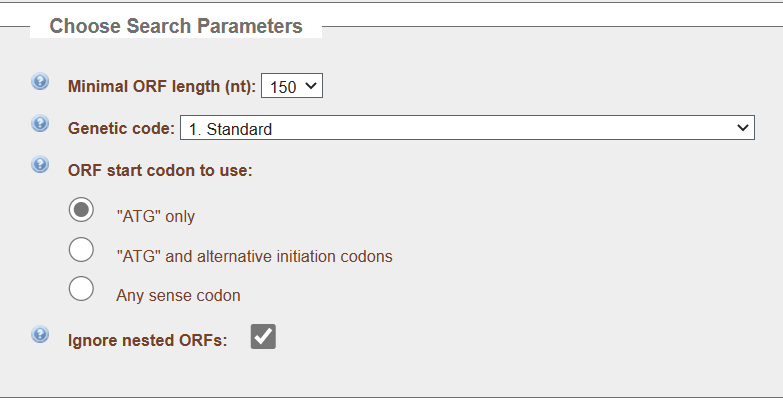
1. Minimal orf length = 75
2. Genetic code = 11 (prokaryotic viruses)
3. ORF start codon = ATG only
4. Ignore nested ORFs = no

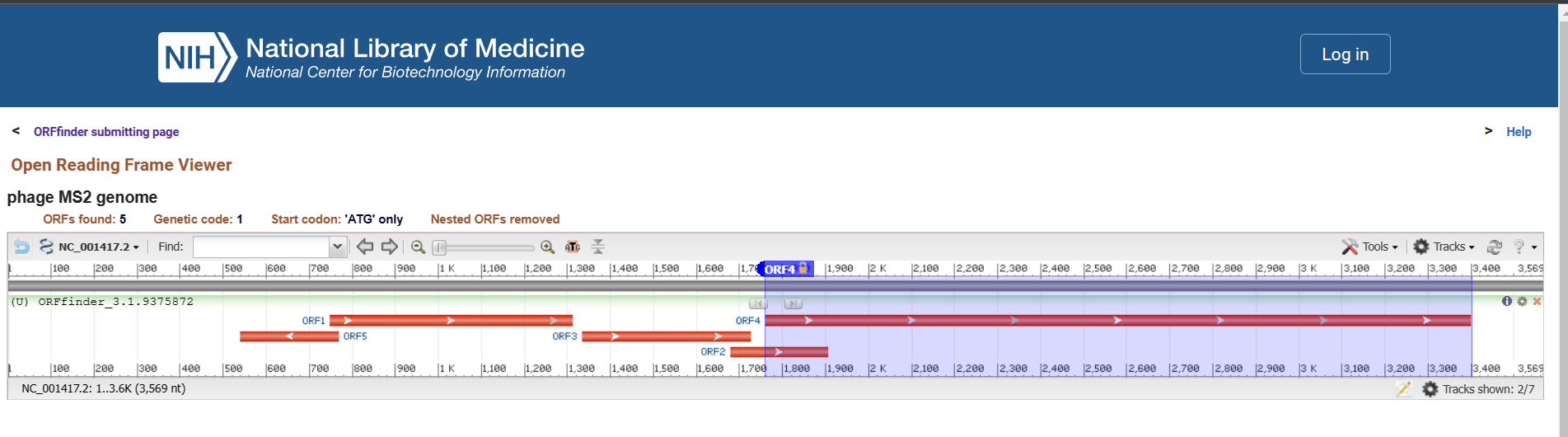


[ORFfinder Viewer - NCBI (nih.gov)](https://www.ncbi.nlm.nih.gov/orffinder/)

Parameters:

1. Minimal orf length = 75
2. Genetic code = 11 (prokaryotic viruses)
3. ORF start codon = ATG only
4. Ignore nested ORFs = no



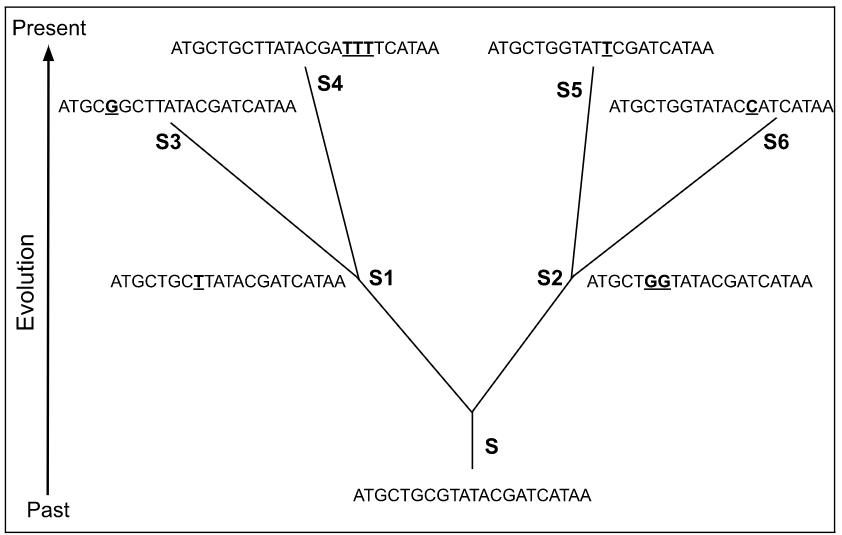


### Using Sequence Databases

1. Using a nucleotide database would be the most appropriate option for finding the required viral genome sequence. NCBI's GenBank, which is a nucleotide database, contains genetic sequences and genomes of various organisms. Protein databases or structure databases would be more useful in finding information about the structures of proteins, nucleic acids, and other complexes. While research publications may also contain viral genome information, nucleotide databases like GenBank are more useful in finding whole genomes and all the relevant information associated with them.
2. The gene can be found online here: [176120924 - Nucleotide Result](https://www.ncbi.nlm.nih.gov/nuccore/176120924)
   1. What is the length of the sequenced genome?  
      3569 bp
   2. What is the Accession (with version) of the sequenced genome?  
      Accession number with version: NC\_001417.2
   3. What are the first 10 bases of the genome sequence?  
      gggtgggacc
   4. Which country did this sample originate in?

### Identify mutations from evolutionary trajectories

The following tree of evolution shows some key type of changes that can happen in a nucleotide sequence during evolution. To learn more about them, you need to answer the following questions.



Instructions:

Fill in the table with the original and changed nucleotide(s), and the type of evolutionary change for each case.

| From species X to Y | Original | Changed | Type of change |
| --- | --- | --- | --- |
| S to S1 | CGT | CTT | Substitution |
| S to S2 | GCG | GG | Deletion |
| S1 to S3 | GCT | GCG | Substitution |
| S1 to S4 | CGATCA | CGATTTTCA | Insertion |
| S2 to S5 | ATA | ATT | Substitution |
| S2 to S6 | CGA | CCA | Substitution |

### Pairwise alignment and analysis

Open the FASTA sequences of Hemoglobin protein chain A from Humans and from Chimpanzees. Perform a pairwise alignment between the two sequences using the Needleman-Wunsch algorithm. How different are the human and Chimpanzee hemoglobins? What and where are the differences?

**Sequence for hemoglobin A chain from humans:**

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG

KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP

AVHASLDKFLASVSTVLTSKYR

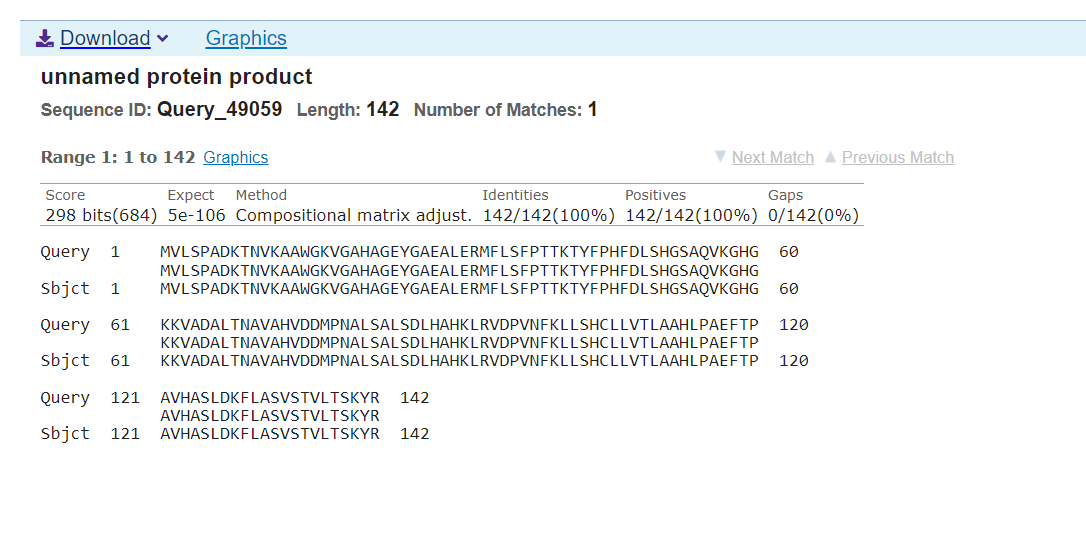
**Sequence for hemoglobin A chain from chimpanzee:**

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG

KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP

AVHASLDKFLASVSTVLTSKYR

**Blast alignment results**



**Results from Needleman-Wunsch algorithm:**

########################################

# Program: needle

# Rundate: Mon 6 Mar 2023 05:11:50

# Commandline: needle

# -auto

# -stdout

# -asequence emboss\_needle-E20230306-051419-0837-1260708-p1m.asequence

# -bsequence emboss\_needle-E20230306-051419-0837-1260708-p1m.bsequence

# -datafile EBLOSUM62

# -gapopen 10.0

# -gapextend 0.5

# -endopen 10.0

# -endextend 0.5

# -aformat3 pair

# -sprotein1

# -sprotein2

# Align\_format: pair

# Report\_file: stdout

########################################

#=======================================

#

# Aligned\_sequences: 2

# 1: EMBOSS\_001

# 2: EMBOSS\_001

# Matrix: EBLOSUM62

# Gap\_penalty: 10.0

# Extend\_penalty: 0.5

#

# Length: 142

# Identity: 142/142 (100.0%)

# Similarity: 142/142 (100.0%)

# Gaps: 0/142 ( 0.0%)

# Score: 733.0

#

#

#=======================================

EMBOSS\_001 1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS 50

||||||||||||||||||||||||||||||||||||||||||||||||||

EMBOSS\_001 1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS 50

EMBOSS\_001 51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK 100

||||||||||||||||||||||||||||||||||||||||||||||||||

EMBOSS\_001 51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK 100

EMBOSS\_001 101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 142

||||||||||||||||||||||||||||||||||||||||||

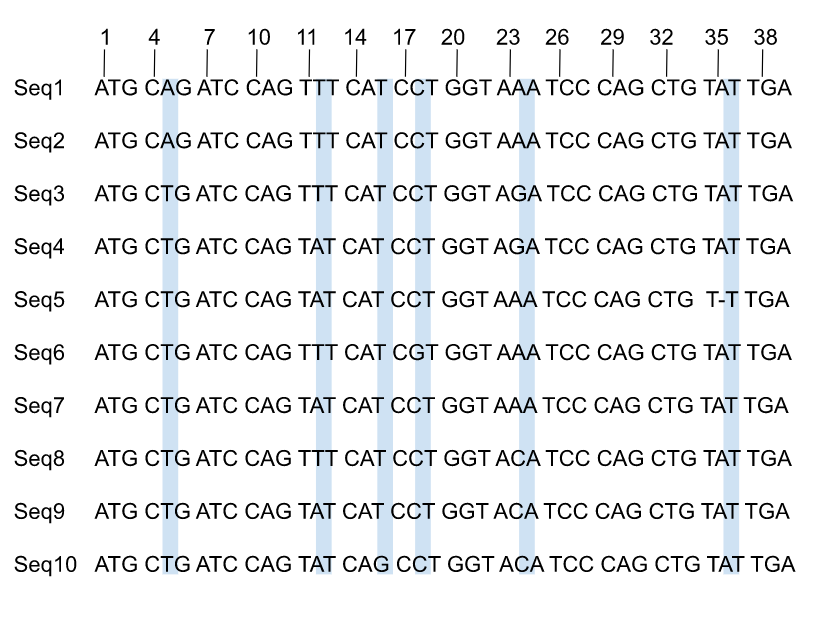
EMBOSS\_001 101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 142

#---------------------------------------

#---------------------------------------

**There is no difference between humans and chimpanzee’s hemoglobin chain A as there is 100% identity.**

### Identify mutations from an MSA

See the following sequence alignment. The positions at which the sequences have some differences are highlighted in light-blue. That means at least some sequences have a mutation at those positions. In which of those cases can you confidently say what may have been the original nucleotide, and which one is the mutated nucleotide? 

Instructions:

* Fill in this table:

| Position | Original nucleotide | Mutated nucleotide |
| --- | --- | --- |
| 5 | T | A |
| 12 | T/A | A/T |
| 16 | T | G |
| 18 | C | G |
| 24 | A | C/G |
| 36 | T | * (deletion) |

**Which one was difficult to guess. Why?**

The position 12 is difficult to guess because according to the parsimony rule, if the letter occurring at the same position has a 50 % chance to the mutated letter then they have equal probability that it can occur in the ancestor or may not occur in the ancestor.

**Identify the mutations in the human Hemoglobin beta chain sequence by comparing it with Gorilla, Chimpanzee and Macaques.**

**Input sequence:**

Gorilla >lcl|NW\_022154674.1\_cds\_XP\_018891709.1\_1 [gene=LOC101126932] [db\_xref=GeneID:101126932] [protein=hemoglobin subunit beta] [protein\_id=XP\_018891709.1] [location=complement(join(872372..872500,873350..873572,873703..873794))] [gbkey=CDS]

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAATGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

Homo sapiens >lcl|NC\_000011.10\_cds\_NP\_000509.1\_1 [gene=HBB] [db\_xref=CCDS:CCDS7753.1,Ensembl:ENSP00000333994.3,GeneID:3043] [protein=hemoglobin subunit beta] [protein\_id=NP\_000509.1] [location=complement(join(5225598..5225726,5226577..5226799,5226930..5227021))] [gbkey=CDS]

ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

Chimpanzee >lcl|NC\_036890.1\_cds\_XP\_508242.1\_1 [gene=HBB] [db\_xref=GeneID:450978] [protein=hemoglobin subunit beta] [protein\_id=XP\_508242.1] [location=complement(join(5086135..5086263,5087114..5087336,5087467..5087558))] [gbkey=CDS]

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

Macaques >lcl|NC\_041767.1\_cds\_NP\_001157900.1\_1 [gene=HBB] [db\_xref=GeneID:715559] [protein=hemoglobin subunit beta] [exception=annotated by transcript or proteomic data] [protein\_id=NP\_001157900.1] [location=join(63067688..63067779,63067910..63068132,63068992..63069120)] [gbkey=CDS]

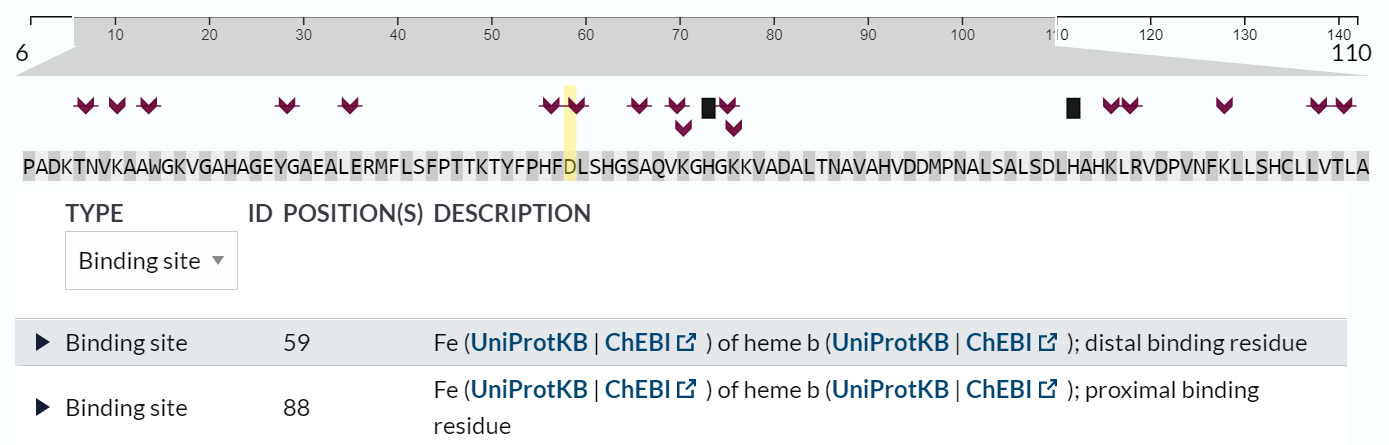
ATGGTGCATCTGACTCCTGAGGAGAAGAATGCCGTCACCACCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCTCTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTTGGTGCCTTTAGTGATGGCCTGAATCACCTGGACAACCTCAAGGGTACCTTTGCCCAGCTCAGTGAGCTGCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCGCAAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTACCACTAA

Results from clustal omega :

<http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?tool=clustalo&jobId=clustalo-E20230306-063956-0354-25482046-p1m>

### Finding functional and structural information from Uniprot

1. Which positions and which amino acids are involved in the binding?



Binding locations: 59 and 88

Amino acids involved:

* At 59th position:

1. What are the amino acids binding to?