Tutorial 3. Heuristic Searching

The Chinese University of Hong Kong

BMEG3102 Bioinformatics

TA: Yizhen Chen

Agenda

- Heuristic searching
  - FASTA
  - BLAST
- FASTA file format
- E-value for BLAST
- Variations of BLAST
- Multiple sequence alignment

Introduction to heuristic alignment

Drawbacks of D.P: Time and space consuming

- To align two sequences of lengths mand n
  - Time complexity: O(*mn*)
  - Space complexity: O(*mn*)

- To find the sequence in a database with **l** length-*n*sequences that is closest to a query sequence of length *m* (Suppose we consider the database sequences one by one)
  - Time complexity: O(**l***mn*)
  - Space complexity: O(*mn*)

- Features of heuristic alignment:
  - Close to (possible to sometimes achieve) optimal solutions
  - Usually faster than optimal methods

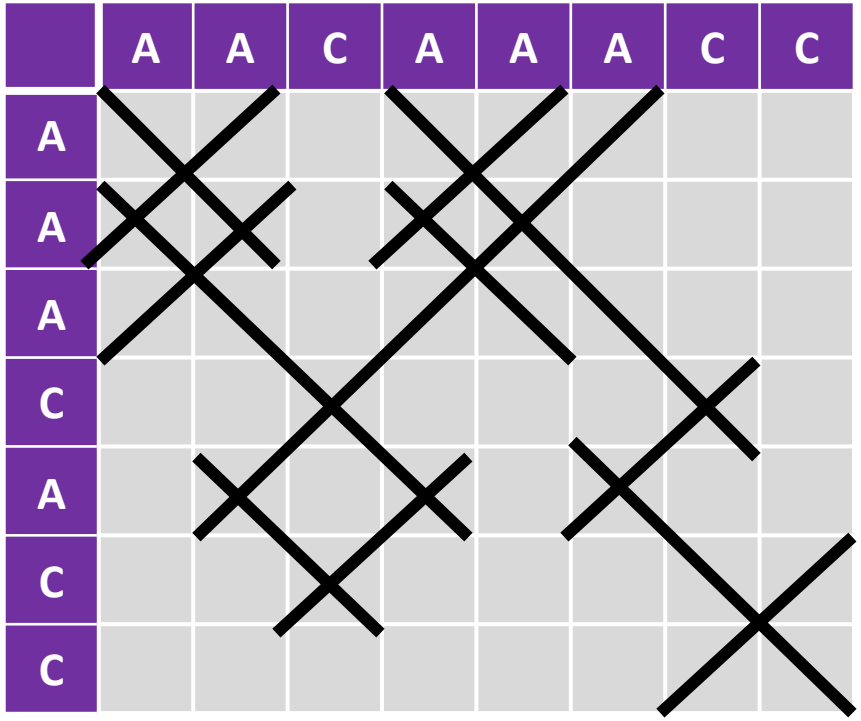
Heuristic alignment

- Basic ideas:
  - Quickly identify regions with high similarity
  - Combine and refine these initial results

- Common methods:
  - FASTA
  - BLAST

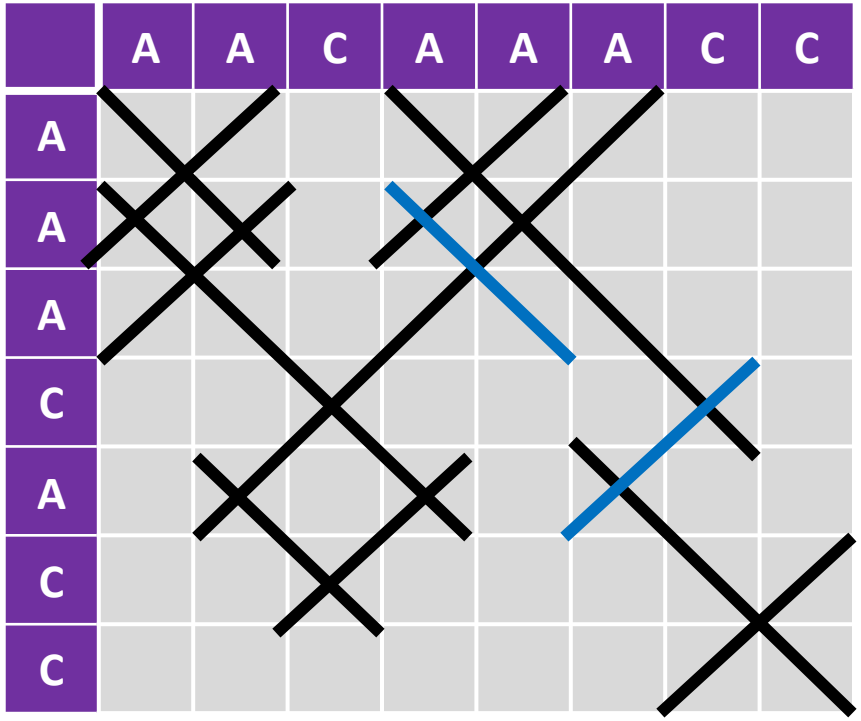
Example: Dot Plot

- A Dot Plot is provided. Answer the questions.



Example: Dot Plot

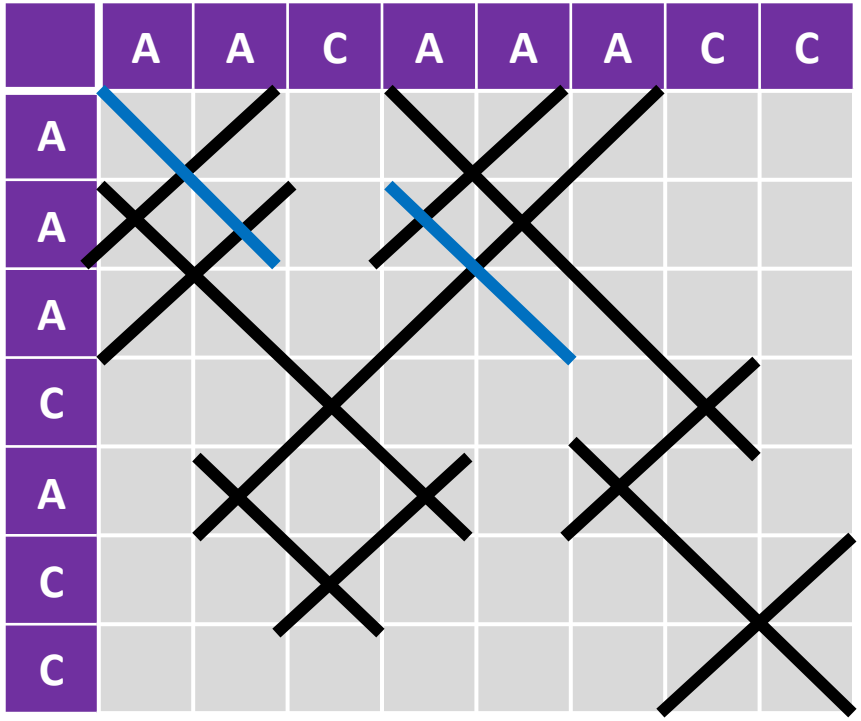
- A Dot Plot is provided. Answer the questions.



- Will you join the two lines in blue?
- No, they are not in the same direction

Example: Dot Plot

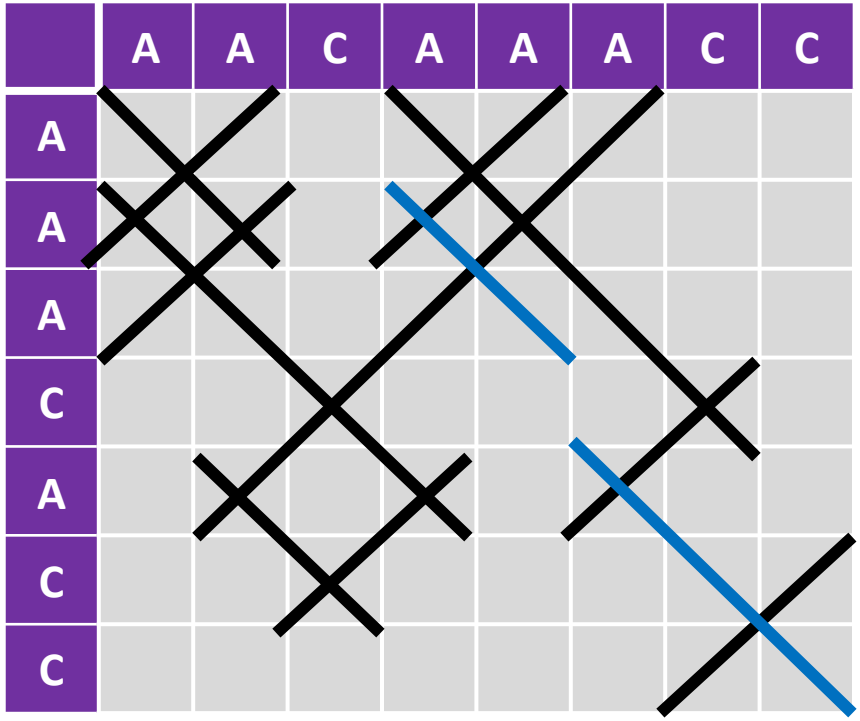
- A Dot Plot is provided. Answer the questions.



- Will you join the two lines in blue?
- No, they share one “A” in one sequence

Example: Dot Plot

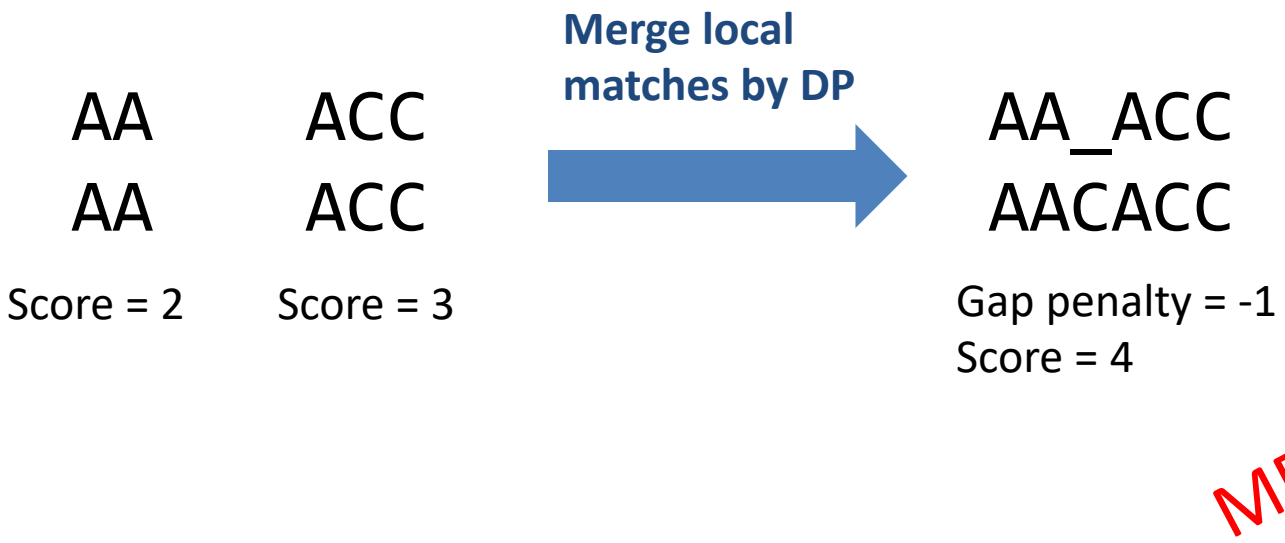
- A Dot Plot is provided. Answer the questions.



- Will you join the two lines in blue?
- It depends, because …

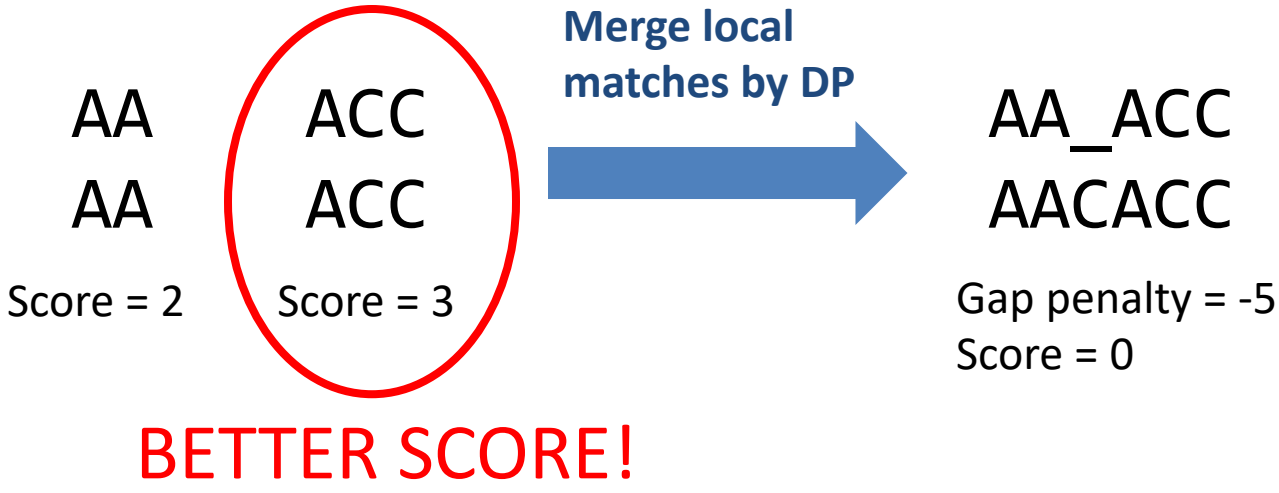
Example: Dot Plot

- It depends on the scoring scheme
- If merging two plots by DP will give a higher alignment score (e.g. match = 1, gap penalty = –1), we will merge the alignments



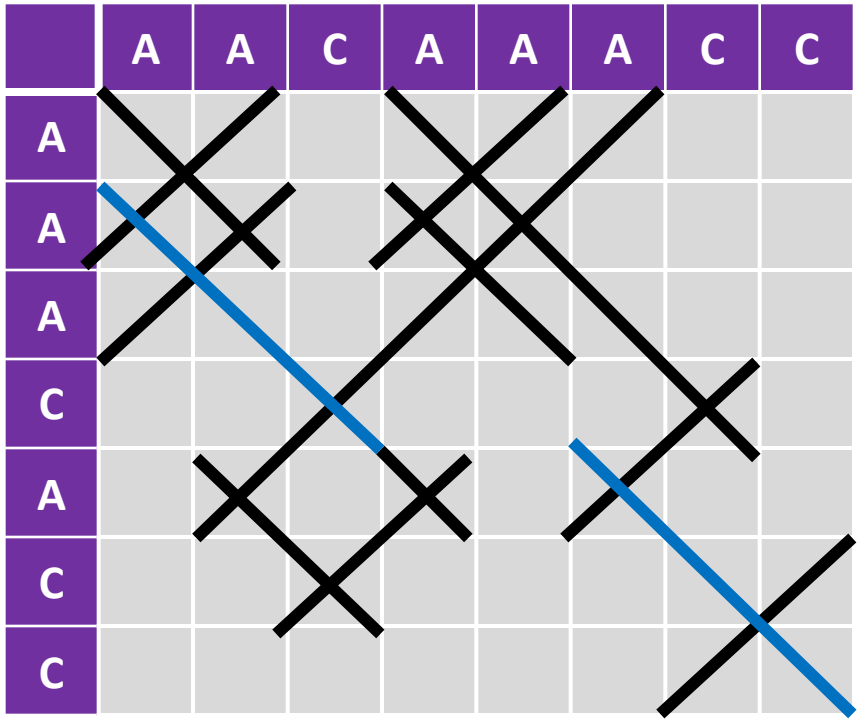
Example: Dot Plot

- It depends on the scoring scheme
- If merging two plots by DP will give a lower alignment score (e.g. match = 1, gap penalty = –5), we will not merge the alignments



Example: Dot Plot

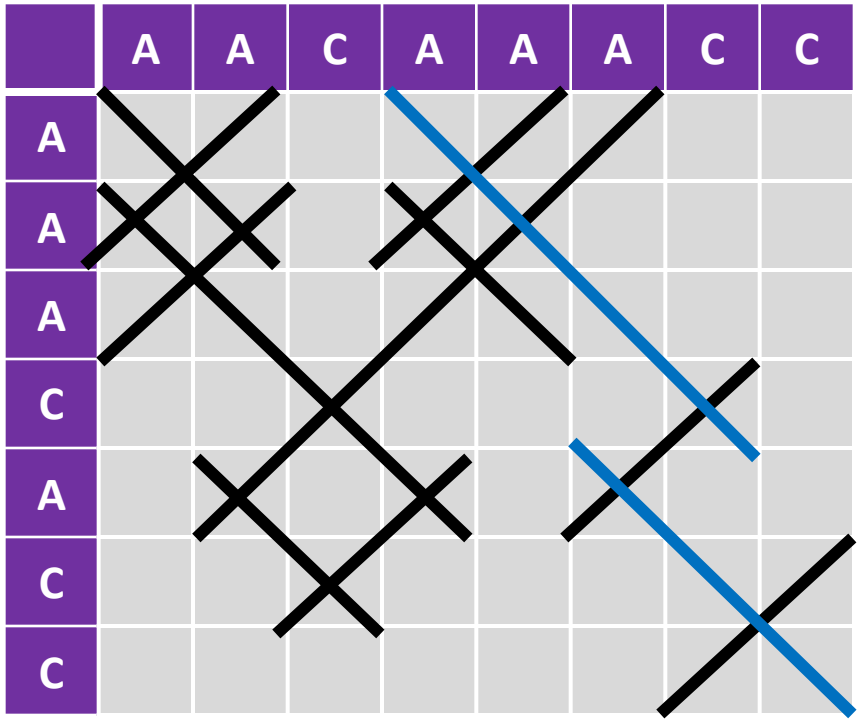
- A Dot Plot is provided. Answer the questions.



- If the blue lines form parts of an alignment that represent the actual evolutionary events, how do you interpret the pattern formed by them?
- Indel

Example: Dot Plot

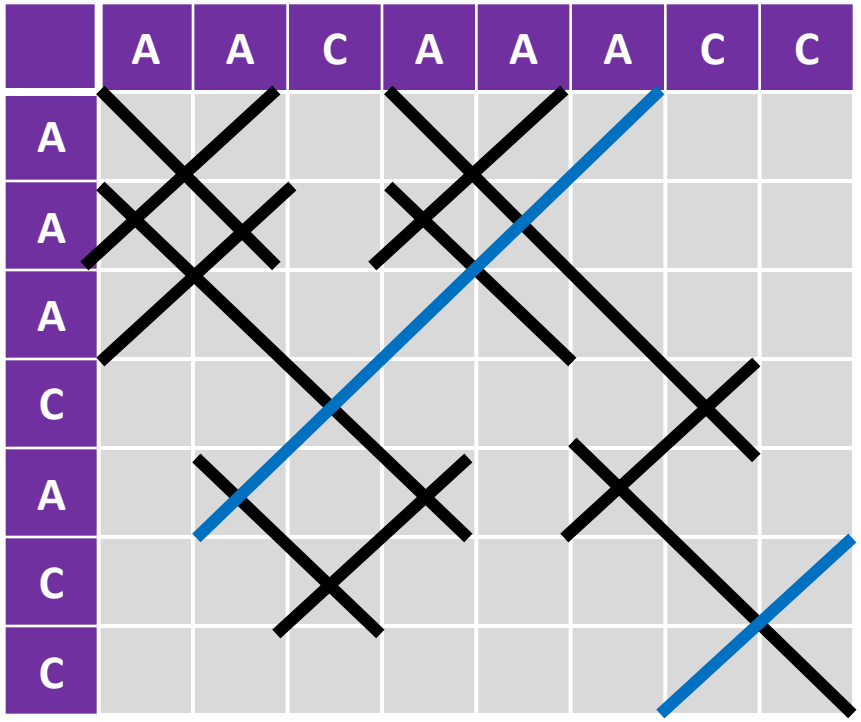
- A Dot Plot is provided. Answer the questions.



- If the blue lines form parts of an alignment that represent the actual evolutionary events, how do you interpret the pattern formed by them?
- Duplication

Example: Dot Plot

- A Dot Plot is provided. Answer the questions.



- If the blue lines form parts of an alignment that represent the actual evolutionary events, how do you interpret the pattern formed by them?
- Inversion and Rearrangement

FASTA: Find local matches

- Construct a lookup table for all *k*-mers (length-*k* subsequences) for a sequence *s*
- E.g., a sequence *s* in database *D*:

```
0         1
12345678901234
CCTACCGATACCGA
```

5-mer in *s*	Position
ACCGA	4, 10
ATACC	8
CCGAT	5
CCTAC	1
CGATA	6
CTACC	2
GATAC	7
TACCG	3, 9

FASTA: Find local matches

- *r*: ACCGATAGC
- *s*: CCTACCGATACCGA

5-mer in *r*	Position
ACCGA	1
ATAGC	5
CCGAT	2
CGATA	3
GATAG	4

5-mer in *s*	Position
ACCGA	4, 10
ATACC	8
CCGAT	5
CCTAC	1
CGATA	6
CTACC	2
GATAC	7
TACCG	3, 9

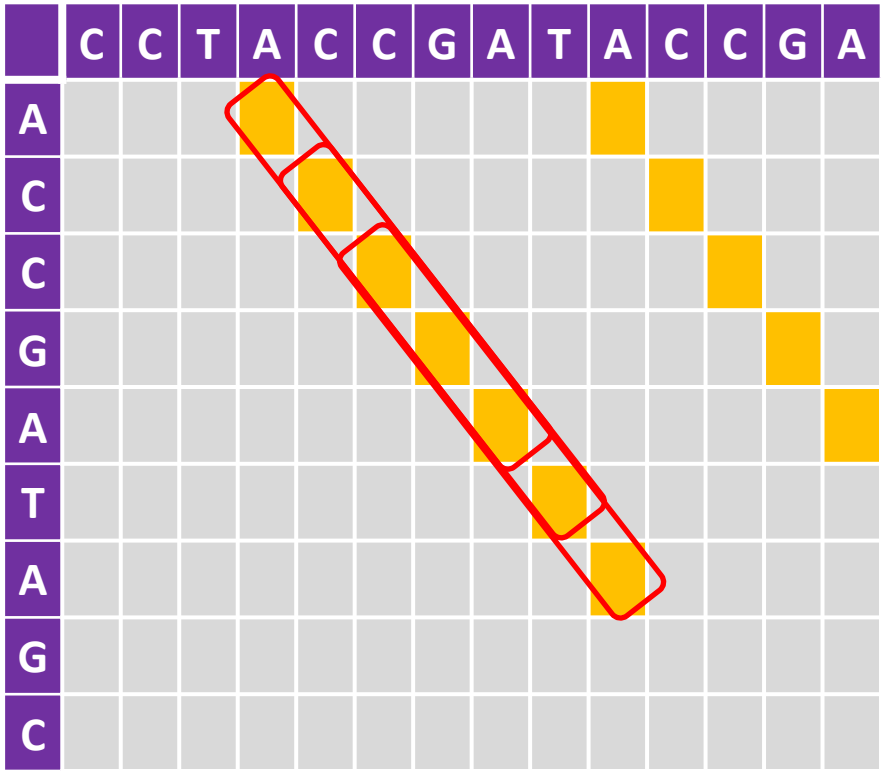
FASTA: Join local matches

- Remember that we do NOT store the dot plot.

- From the lookup table:

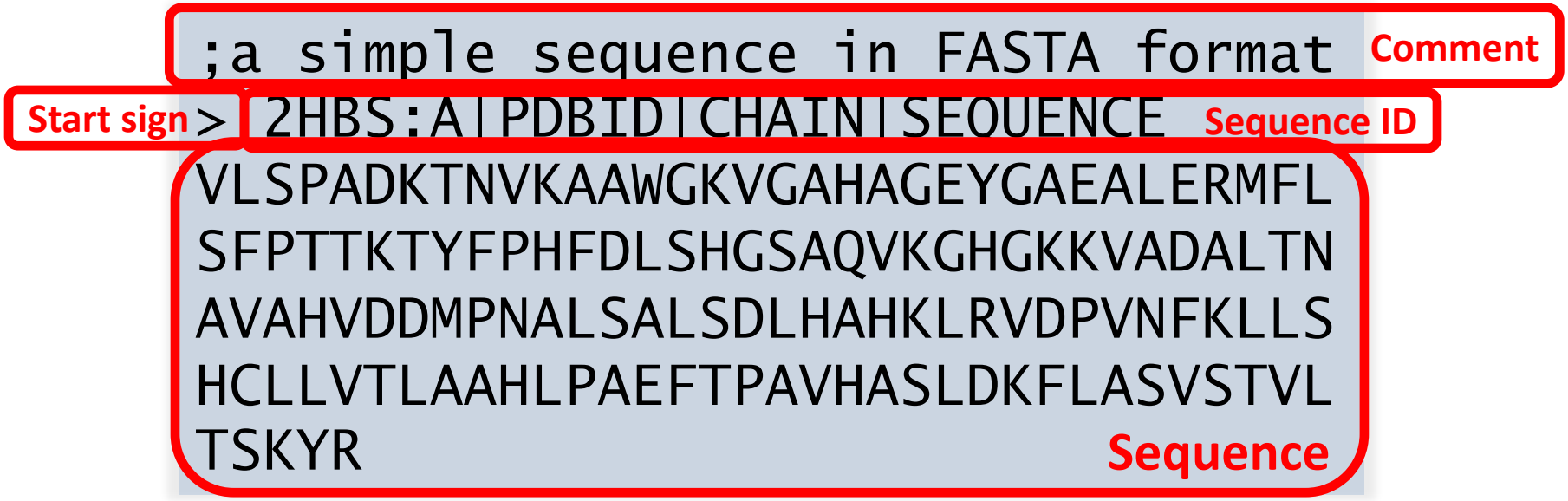
$r[1..5]=s[4..8]=s[10..14]$
$r[2..6]=s[5..9]$
$r[3..7]=s[6..10]$

$r[1..7]=s[4..10]$



FASTA file format

- Save either nucleotide sequences or peptide sequences where nucleotides or amino acids are stored in single letters.



- Here are some more common formats:
- https://www.ebi.ac.uk/ena/submit/data-formats

BLAST

- Find local matches: allow high-scoring inexact matches
- Extend and join local matches
- Evaluate the statistical significance (E-value)

Example: BLAST

- Construct a lookup table for all 5-mer subsequences (sorted in alphabetical order):
  - *r*: CCTACCGATACCGA
- If 1 mismatch is allowed, write down the result of each 5-mer on the query sequence:
  - *s*: ACCTATAGC
- Extend local matches to form longer matches
- Combine matches to give alignments
- Calculate scores, assume that mismatch score = -1, match score = 1

Answer: BLAST

- *r*: CCTACCGATACCGA
- *s*: ACCTATAGC

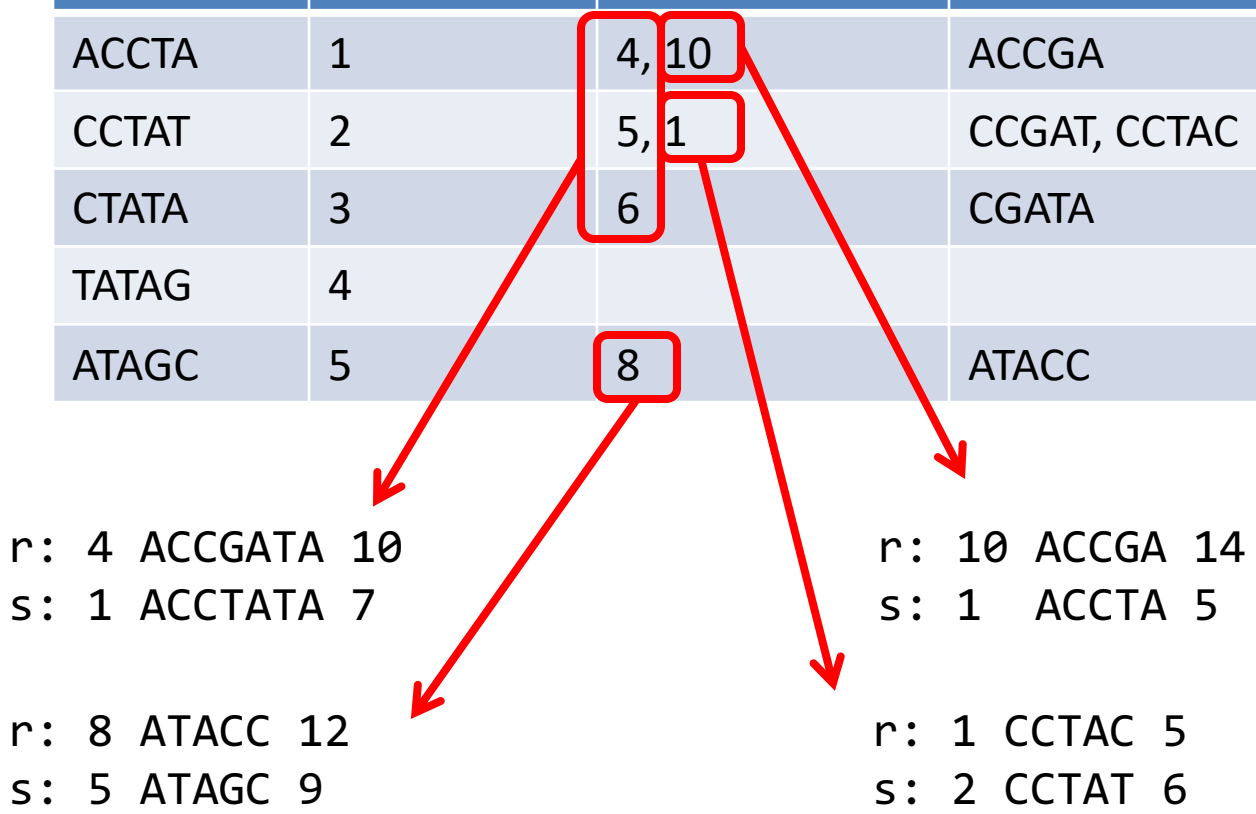
5-mer in *r*	Position
ACCGA	4, 10
ATACC	8
CCGAT	5
CCTAC	1
CGATA	6
CTACC	2
GATAC	7
TACCG	3, 9

5-mer in *s*	Position in *s*	Corr. Match in *r*	Matched 5-mer in *r*
ACCTA	1	4, 10	ACCGA
CCTAT	2	5, 1	CCGAT, CCTAC
CTATA	3	6	CGATA
TATAG	4		
ATAGC	5	8	ATACC

Answer: BLAST

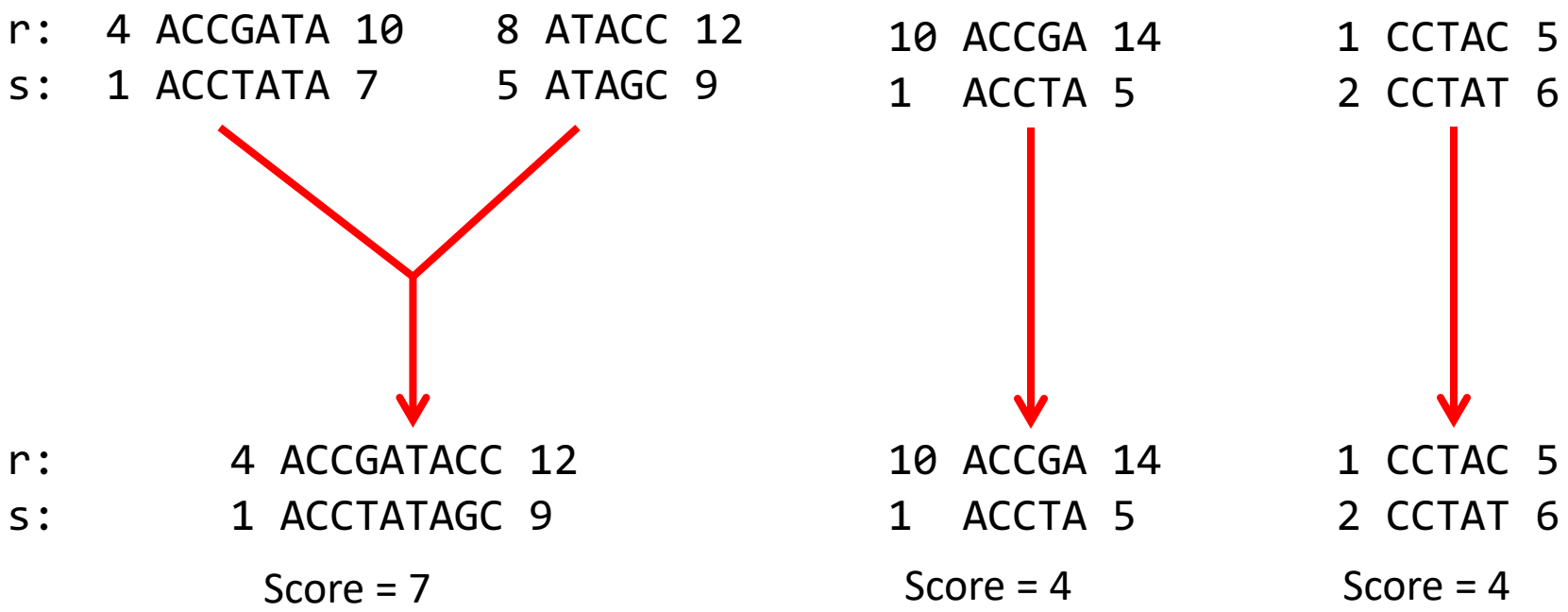
- Extend local matches to longer matches:

5-mer in *s*	Position in *s*	Corr. Match in *r*	Matched 5-mer in *r*
ACCTA	1	4, 10	ACCGA
CCTAT	2	5, 1	CCGAT, CCTAC
CTATA	3	6	CGATA
TATAG	4		
ATAGC	5	8	ATACC



Answer: BLAST

- Combine matches and calculate scores:



BLAST: E-value

- Suppose the query sequence *r* has length *m*, a sequence *s* in the database has length *n*, and a match has score *Q*.
- E-value is the expected (i.e., mean) number of matches with score *Q* or larger when a random sequence of length *m* is compared to a set of random sequences with the same total number and length distribution as the sequences in the database.

- For more information:
- https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect

E-value: An example

- Suppose we are searching “CA” in a database containing ALL RNA sequences of length 4, and get the following result:
  - *r*: `CA`
  - *s:* `CAAC`
- Match: +1; mismatch/indel: -1. So the local alignment score is 2 (exact match).
- Calculate E-value of this alignment.

E-value: An example

- Idea of calculating E-score:
  - Find all possible alignments with score ≥ 2 (in this case, only need to consider the exact match).
  - Divided by all possible combinations of a random sequence with length 2 and a database containing all RNA sequences with length 4.
  - We can consider the following two cases:
    - Case I: the query sequence has two different characters, e.g., GC
    - Case II: the query sequence has the same character, e.g., AA

E-value: An example

- Case I:
  - The query sequence has two different characters
  - Number of possible query sequences: 12
  - Take “GC” as an example:

	Possible *s* in the database	Number of matches in each pair	Number of pairs
2 “GC” in *s*	5’-GCGC-3’	2	1
1 “GC” in *s*	5’-GCXX-3’	1	15
	5’-XGCX-3’		16
	5’-XXGC-3’		15

E-value: An example

- Case II:
  - The query sequence has the same character
  - Number of possible query sequences: 4
  - Take “AA” as an example:

	Possible *s* in the database	Number of matches in each pair	Number of pairs
4 “A” in *s*	5’-AAAA-3’	3	1
3 “A” in *s*	5’-AAAX-3’	2	3
	5’-XAAA-3’		3
2 “A” in *s*	5’-AAXX-3’	1	9
	5’-XAAX-3’		9
	5’-XXAA-3’		9

E-value: An example

- Total number of possible combinations:
  - $4^2 * 4^4 = 4096$
- So the E-value is:

$$\frac{number\ of\ matches\ with\ score \geq 2\ (exact\ match\ in\ this\ case)}{total\ number\ of\ possible\ combinations}$$

$$= \frac{12a+4b}{4096}$$

$$= \frac{12\times(2\times1+1\times(15+16+15))+4\times(3\times1+2\times(3+3)+1\times(9+9+9))}{4096}$$

$$= 0.18164$$

Comparison between FASTA and BLAST

	FASTA	BLAST
Full name	**FAST** for **A**ll kinds of sequences	**B**asic **L**ocal **A**lignment **S**earch **T**ool
Similarities	• Use the idea of dot plot • Use *k*-mers to construct the lookup table • Merge local matches • Join local matches using dynamic programming	
Differences	• Use exact matches • Start with smaller *k* (usually 6-8 for DNA) • Slower in most cases	• Allow mismatches • Start with larger *k* (usually 11 for DNA) • Usually faster • Evaluate the statistical significance

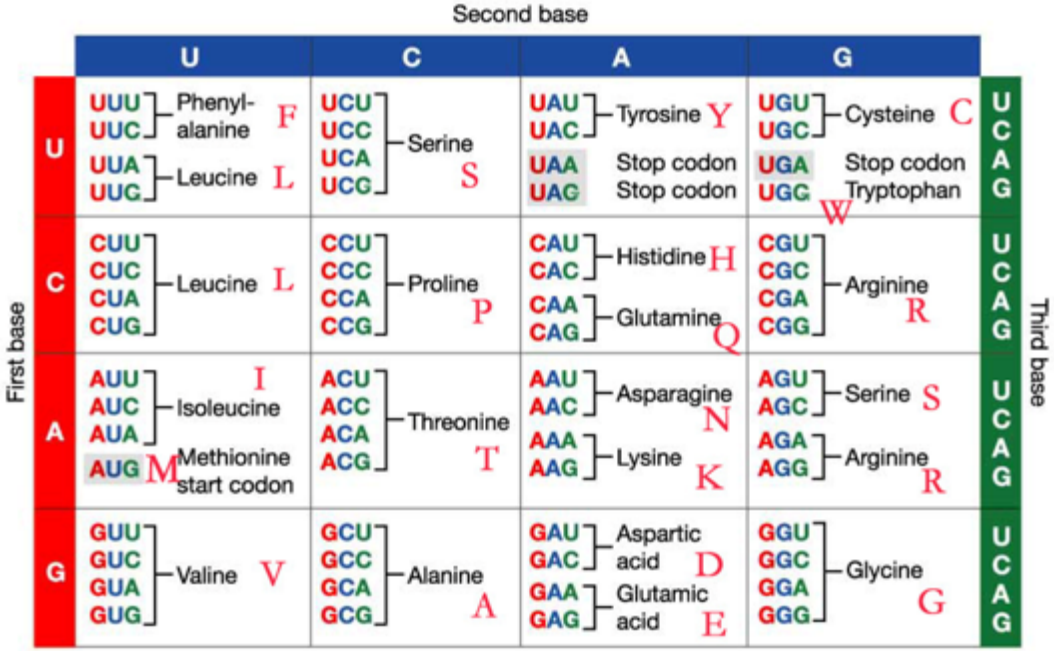
Variations of BLAST

Tool	Query sequence	Database sequences	Comparison
blastn	Nucleotide	Nucleotide	Nucleotide-nucleotide
blastp	Protein	Protein	Protein-protein
blastx	Nucleotide	Protein	6FT-protein
tblastn	Protein	Nucleotide	Protein-6FT
tblastx	Nucleotide	Nucleotide	6FT-6FT

- Why do we translate nucleotide, instead of protein in blastx and **tblastn**?

Recall the codon table

- There are 61 (= 4^3-3 (stop codons)) codons encoding 20 amino acids in total
- On average, 1 amino acid corresponds to 3.05 (= 61/20) codons



Imagine if we translate protein …

- If we translate the protein sequence PV into DNA:
  - P can be {CCA, CCC, CCG, CCT}
  - V can be {GTA, GTC, GTG, GTT}

- Together, PV can be
  - {CCAGTA, CCAGTC, CCAGTG, CCAGTT,
    CCCGTA, CCCGTC, CCCGTG, CCCGTT,
    CCGGTA, CCGGTC, CCGGTG, CCGGTT,
    CCTGTA, CCTGTC, CCTGTG, CCTGTT}

Imagine if we translate protein …

- In the previous example, PV is length 2
- We have 16 (= 4^2) different DNA sequences

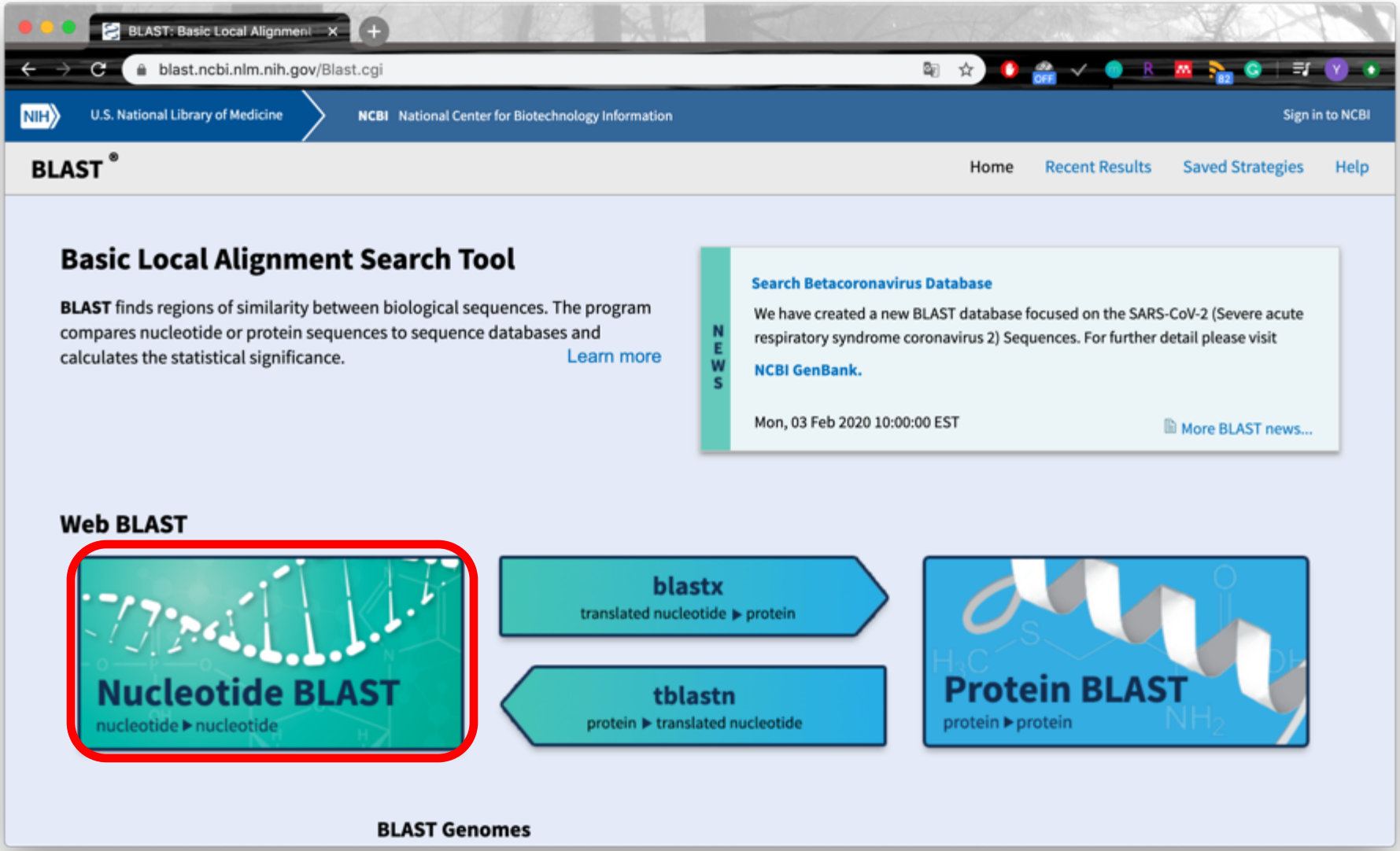
- In general, if we have a protein sequence of length n, there are 3.05^n different possible DNA sequences on average!

Demonstration: blastn

- You are given the following sequence from a human (Homo sapiens) gene:
  - `ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCC GTTACT`

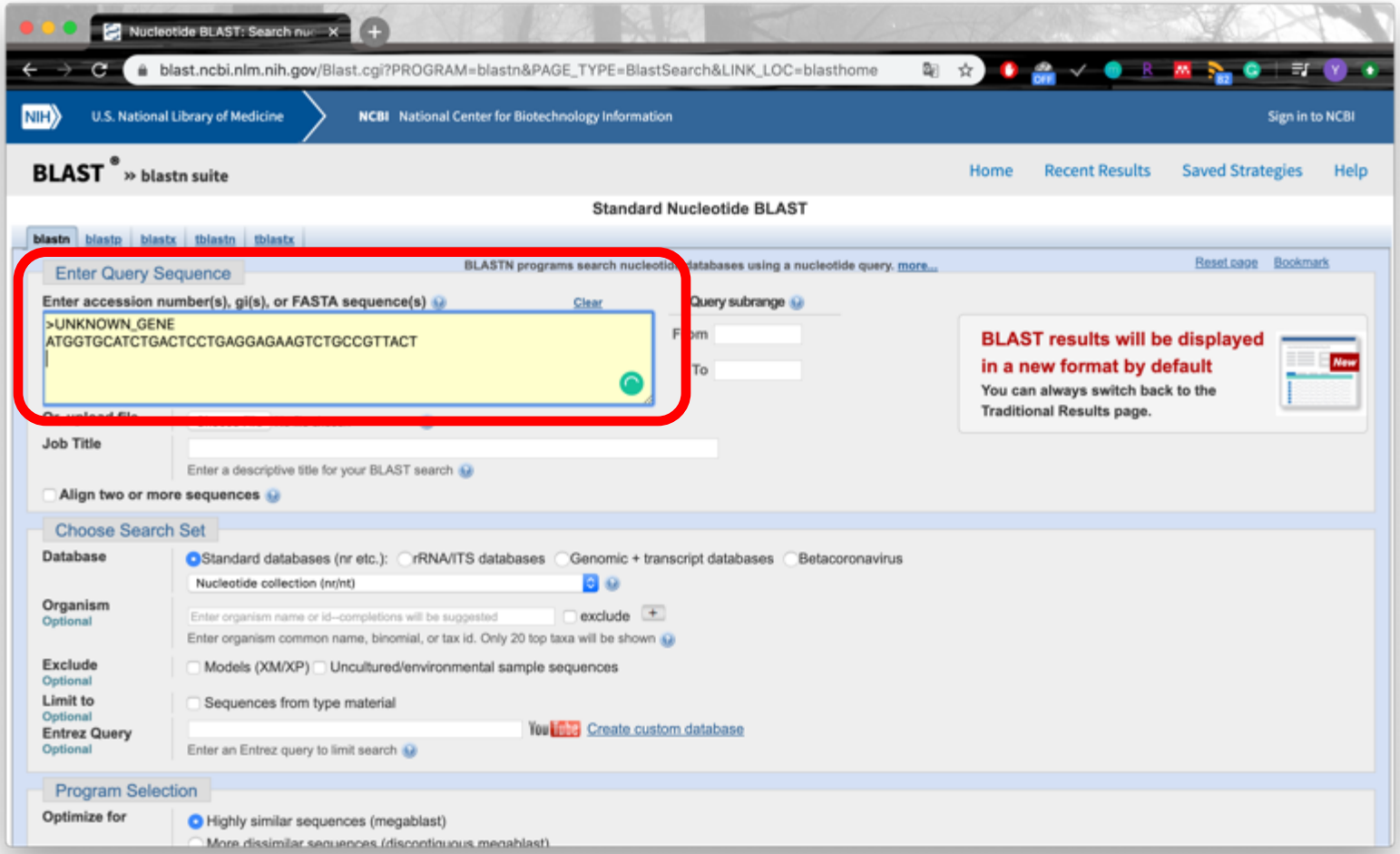
- Go to BLAST, find out the name of the gene that the given sequence most likely comes from.

Step 1: Go to BLAST and Select BLASTN

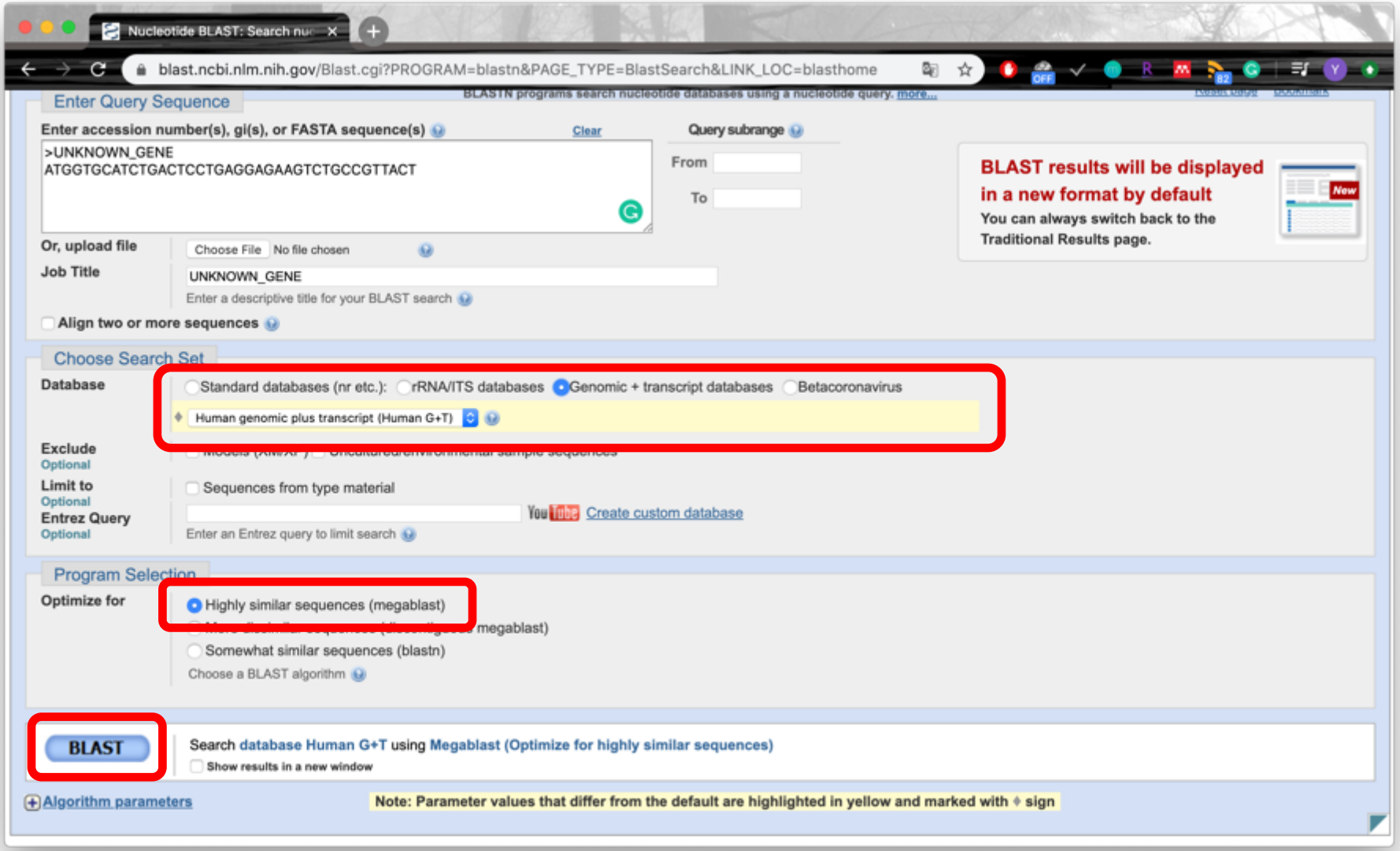


Step 2: Input the sequence in FASTA format

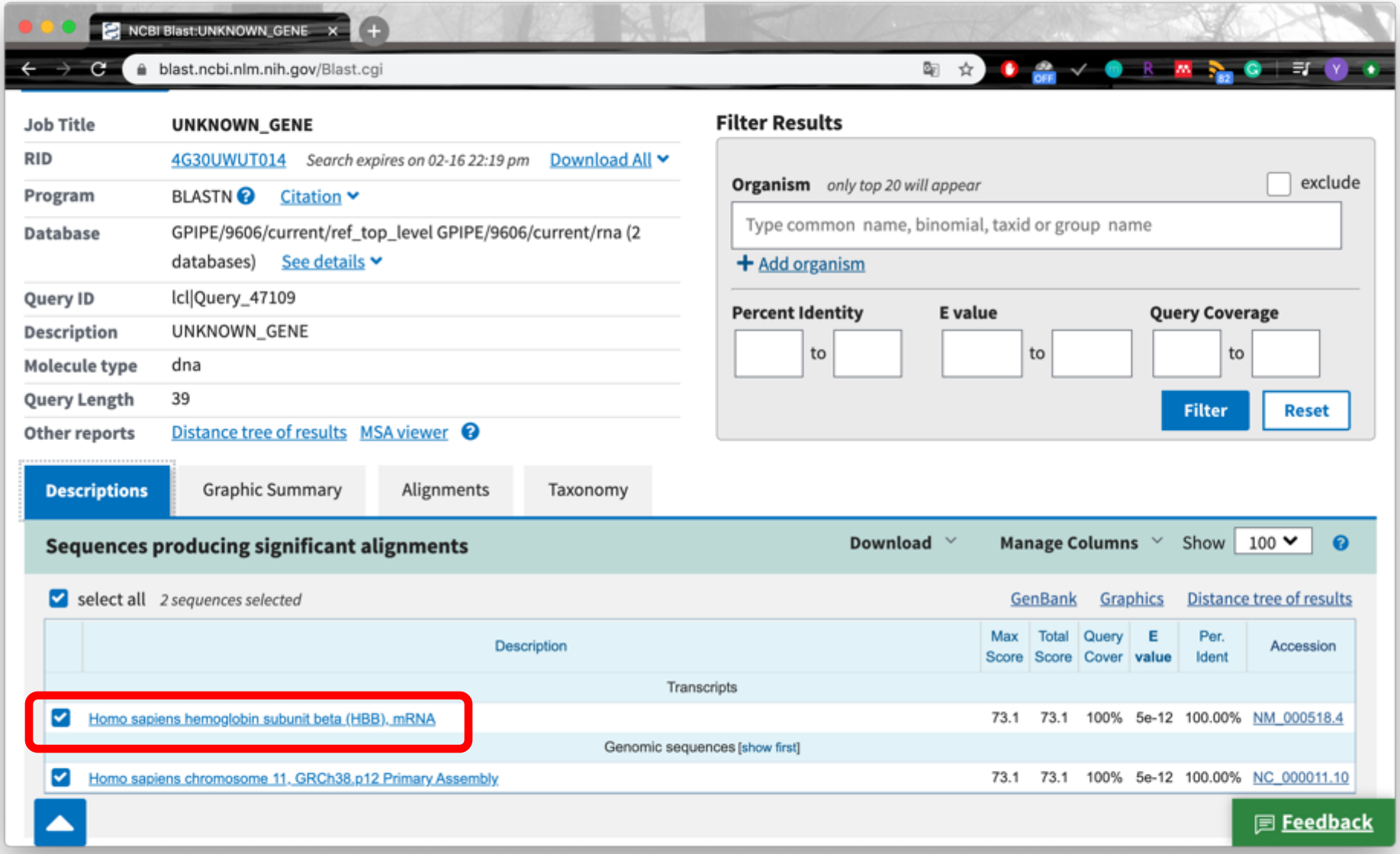
>UNKNOWN_GENE
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT



Step 3: Set the parameters



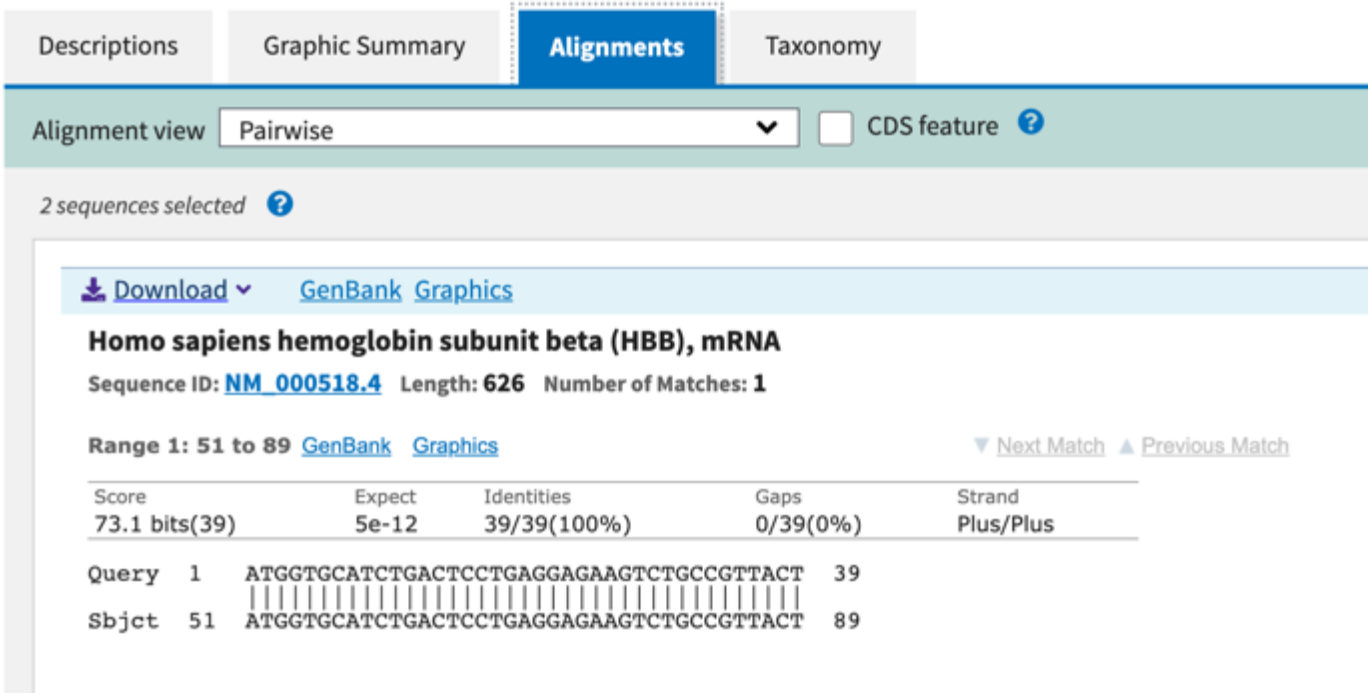
Step 4: Show the results



Step 4: Show the results

Gene: Homo sapiens hemoglobin, beta (HBB), mRNA
E-value: 5e-12
Alignment:

```
Query 1   ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT 39
          |||||||||||||||||||||||||||||||||||||||
Sbjct 51  ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT 89
```



Multiple sequence alignment (MSA)

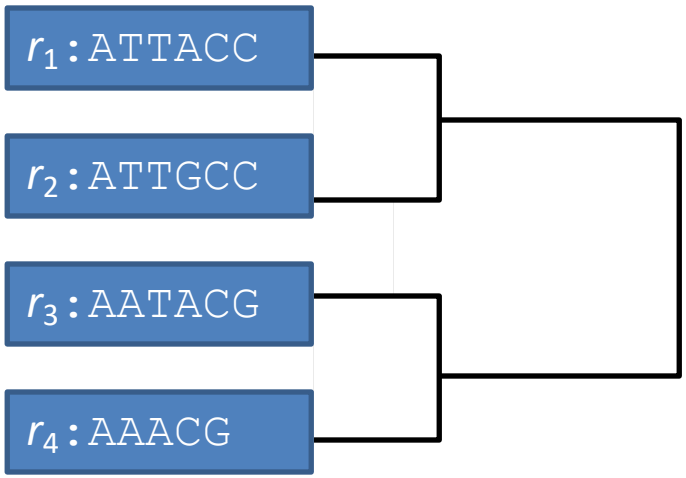
- Evaluate the goodness of an MSA
  - Compare all pairs and calculate the average score
  - Compare each sequence with a consensus sequence and calculate the average score

- Perform good MSA
  - Many different methods
  - One popular progressive alignment method is Clustal family:
    - Construct a tree that captures the distance relationship between the sequences
    - Progressively align the sequences based on the tree

MSA: An example

- Perform MSA for the four sequences below using Clustal method:
  - r_1: ATTACC
  - r_2: ATTGCC
  - r_3: AATACG
  - r_4: AAACG

MSA: An example



One possible tree (detailed procedures of constructing the tree will be covered later)

Alignment order:

1. r_1 vs. r_2
2. r_3 vs. r_4
3. (r_1, r_2) vs. (r_3, r_4)

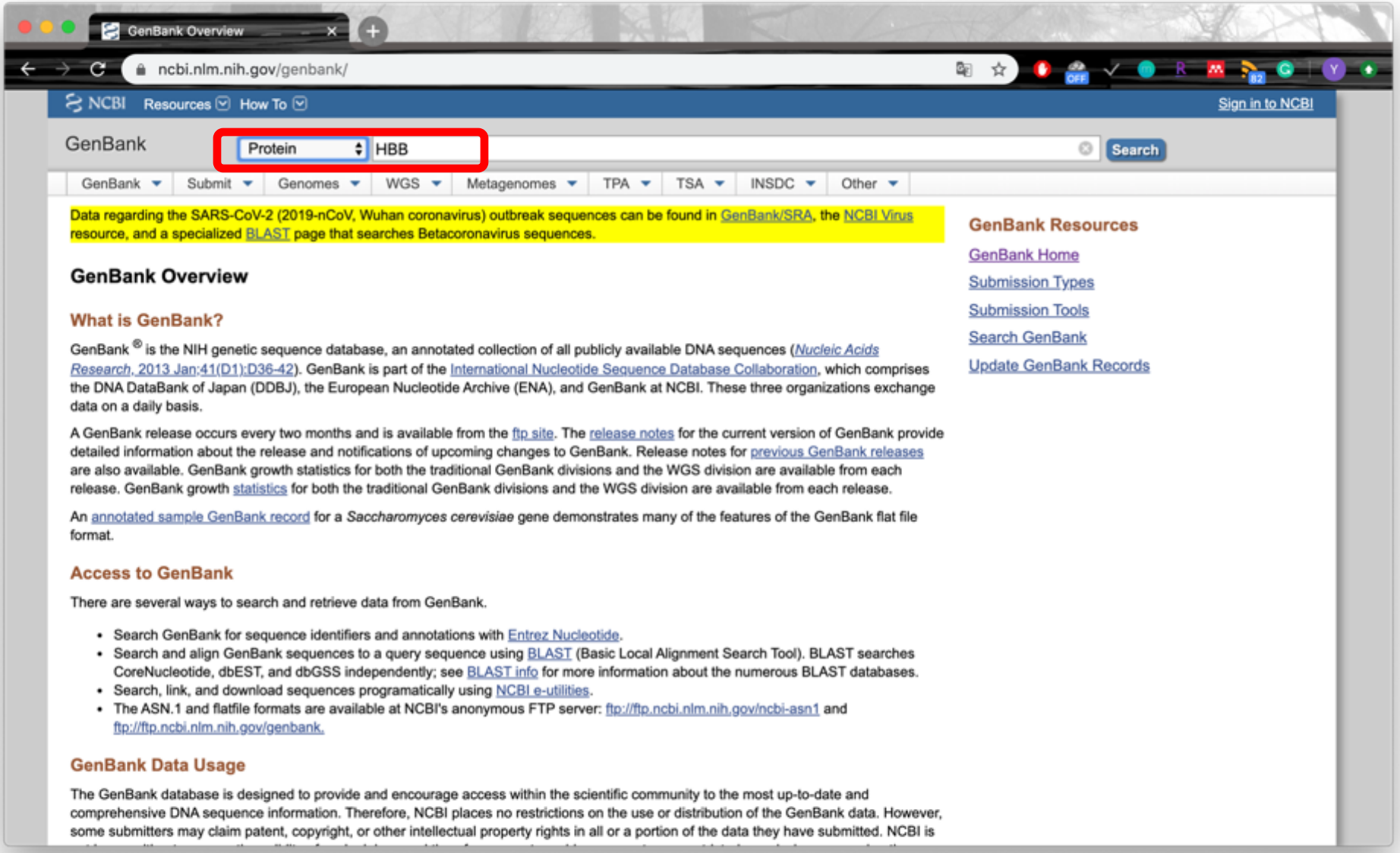
Possible alignment:

r_1: ATTACC
r_2: ATTGCC
r_3: AATACG
r_4: AA_ACG

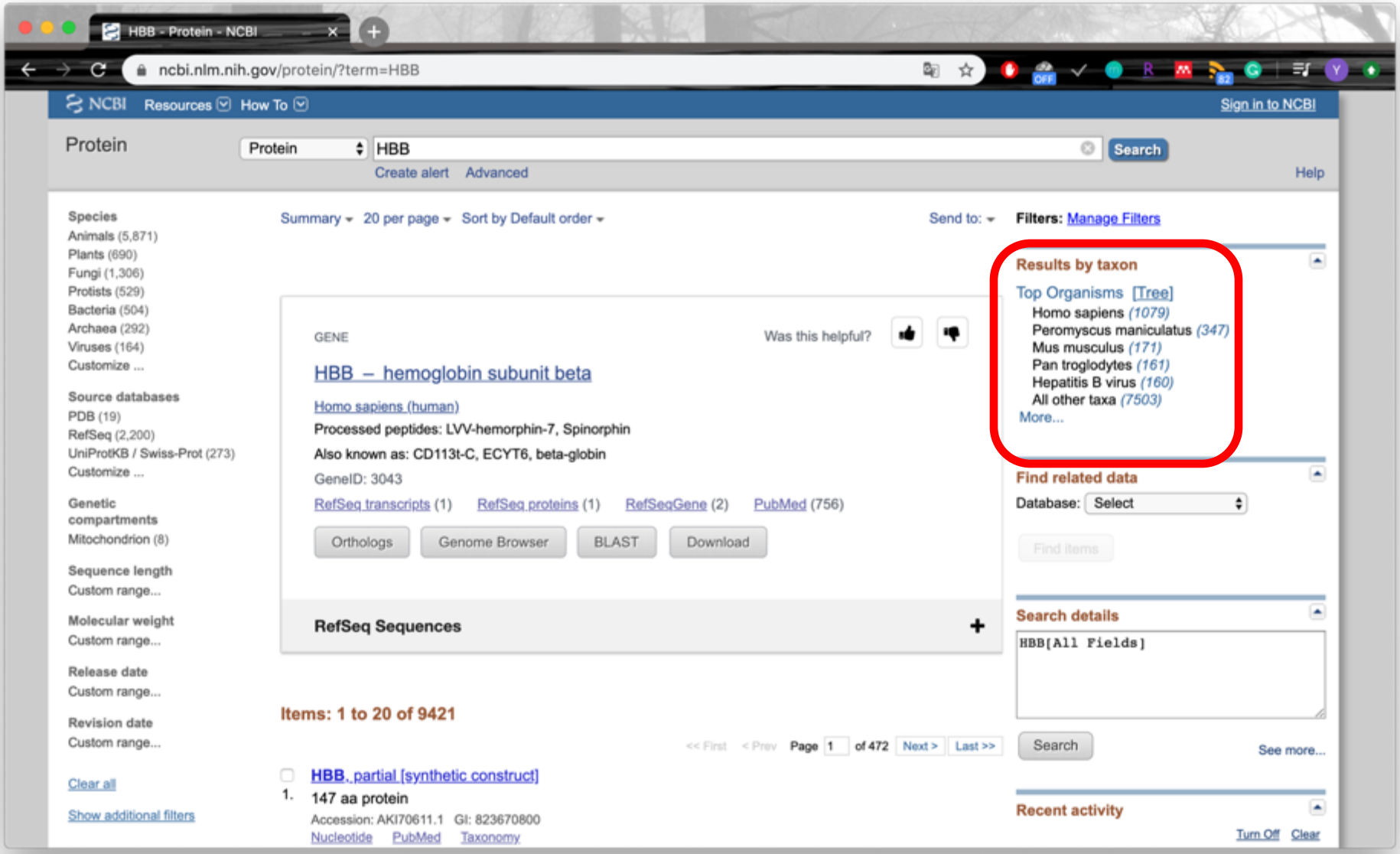
Demonstration: Use a Web tool to perform MSA

- Get the protein sequences of “hemoglobin, beta” (HBB) gene of different organisms from GenBank
- Perform multiple sequence alignment on those sequences

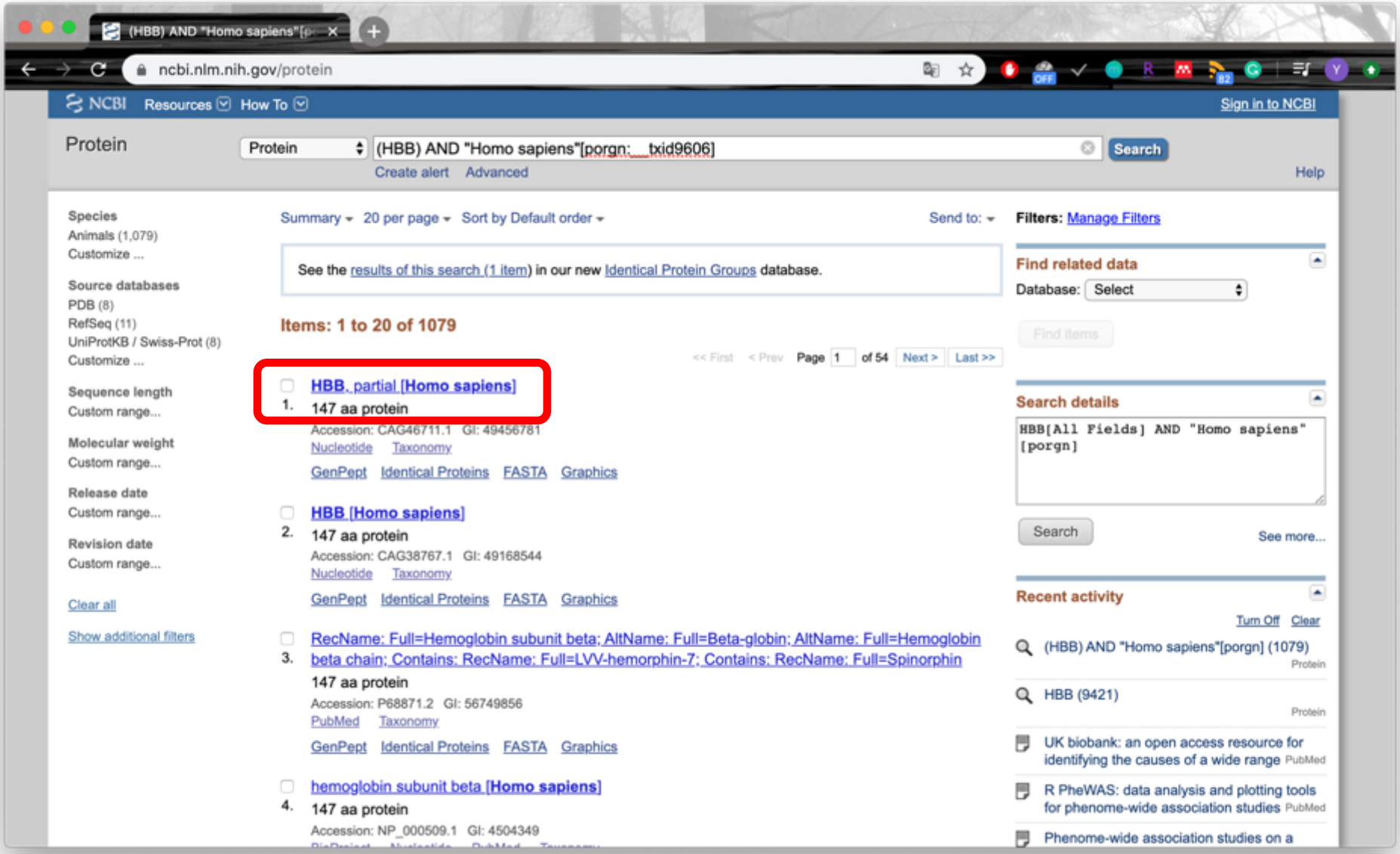
Go to GenBank and search “HBB”



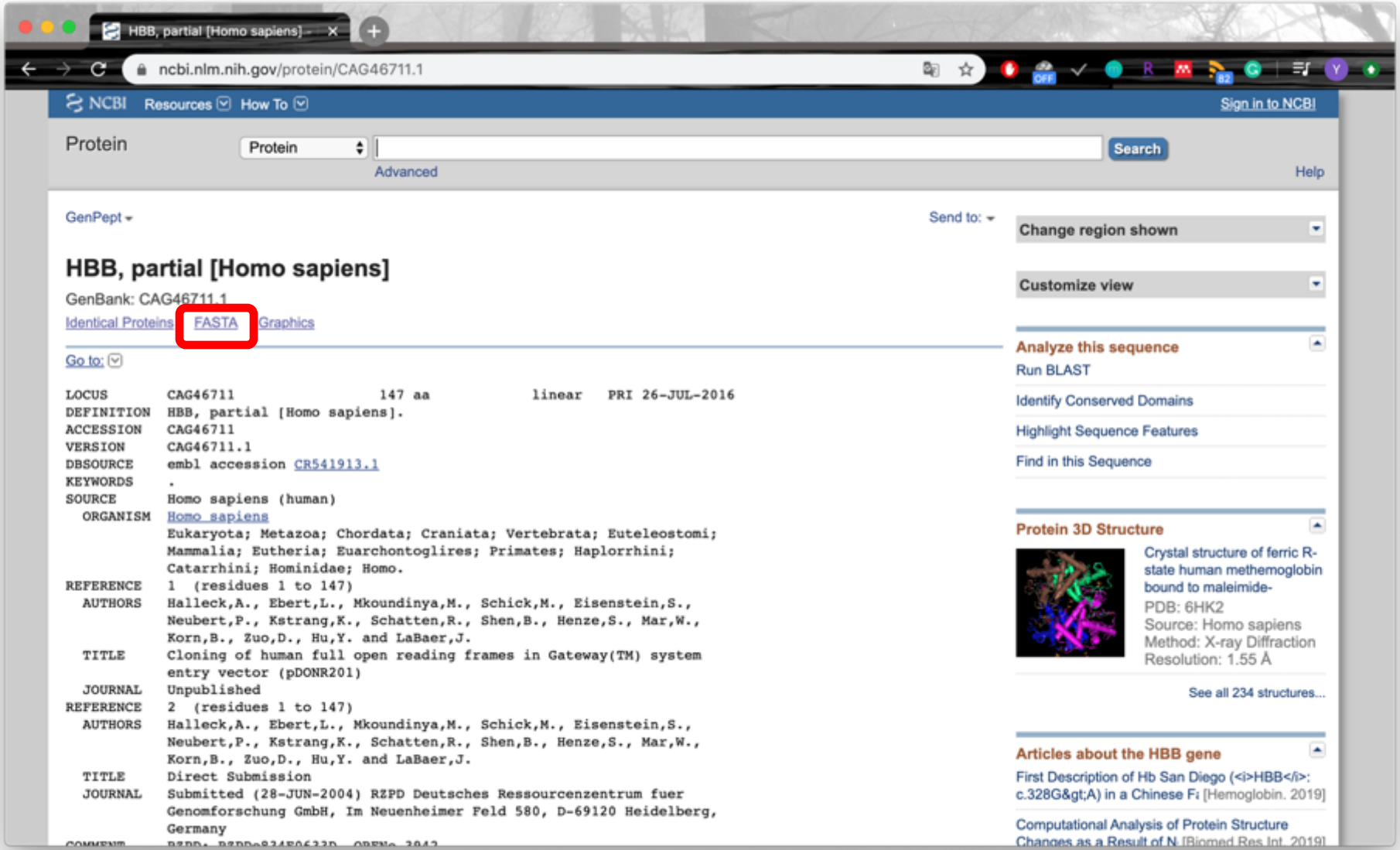
Get HBB protein sequences from GenBank



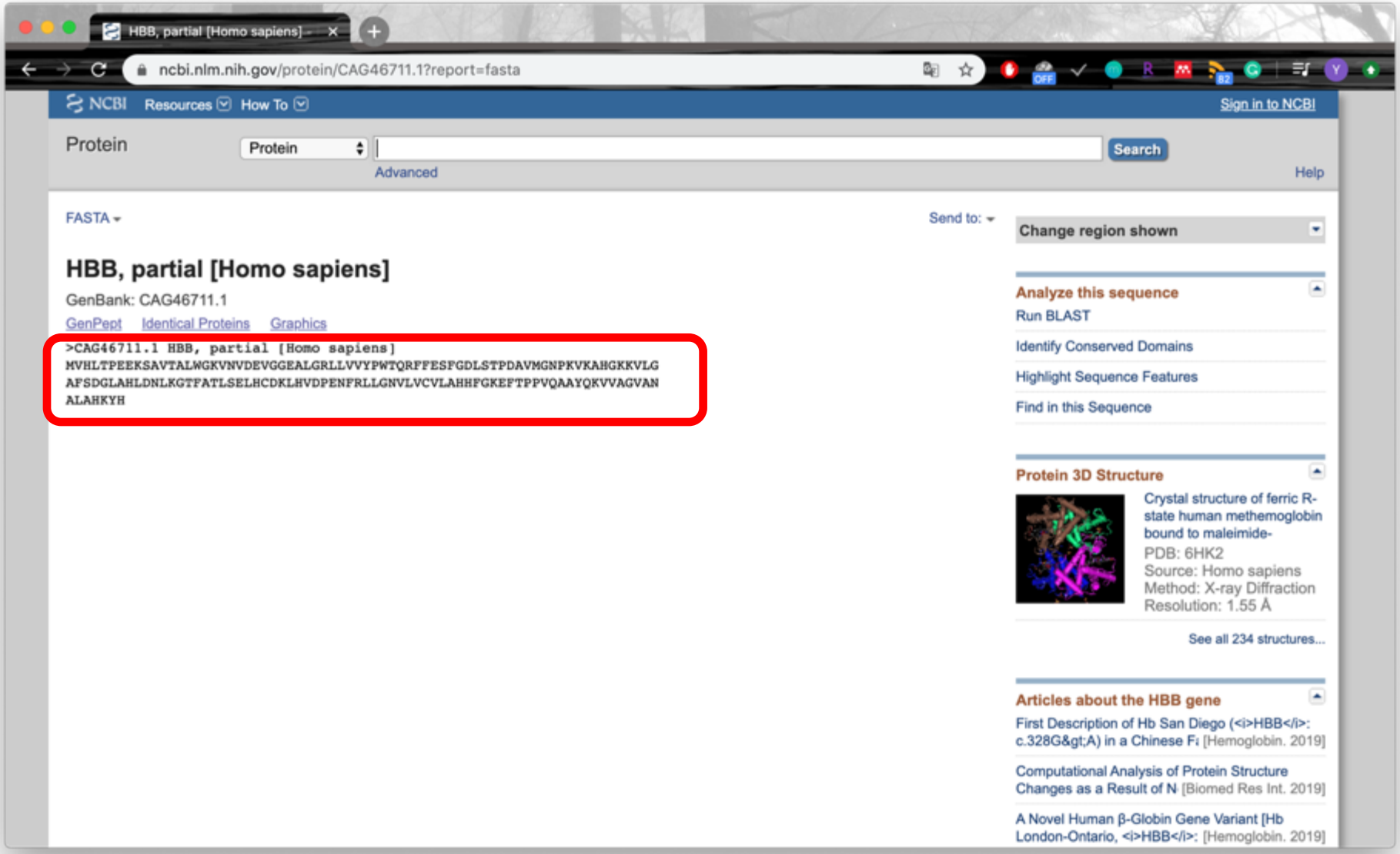
Take Homo Sapiens as an example …



Take Homo Sapiens as an example …



Take Homo Sapiens as an example …



HBB protein sequences from different species

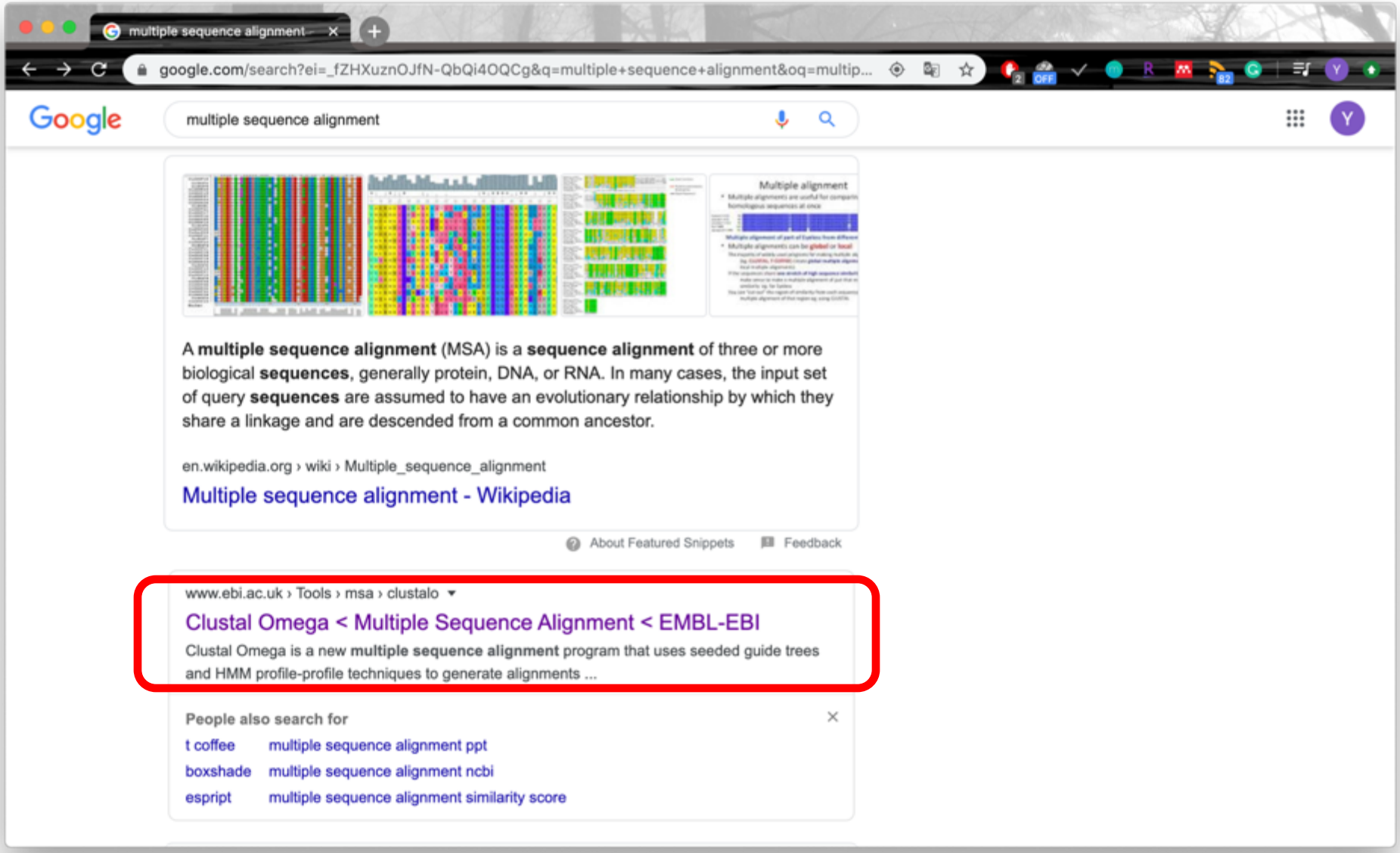
>**Human|**CAG46711.1 HBB, partial [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

>**Chimpanzee|**ACY75418.1 beta globin, partial [Pan troglodytes verus]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR

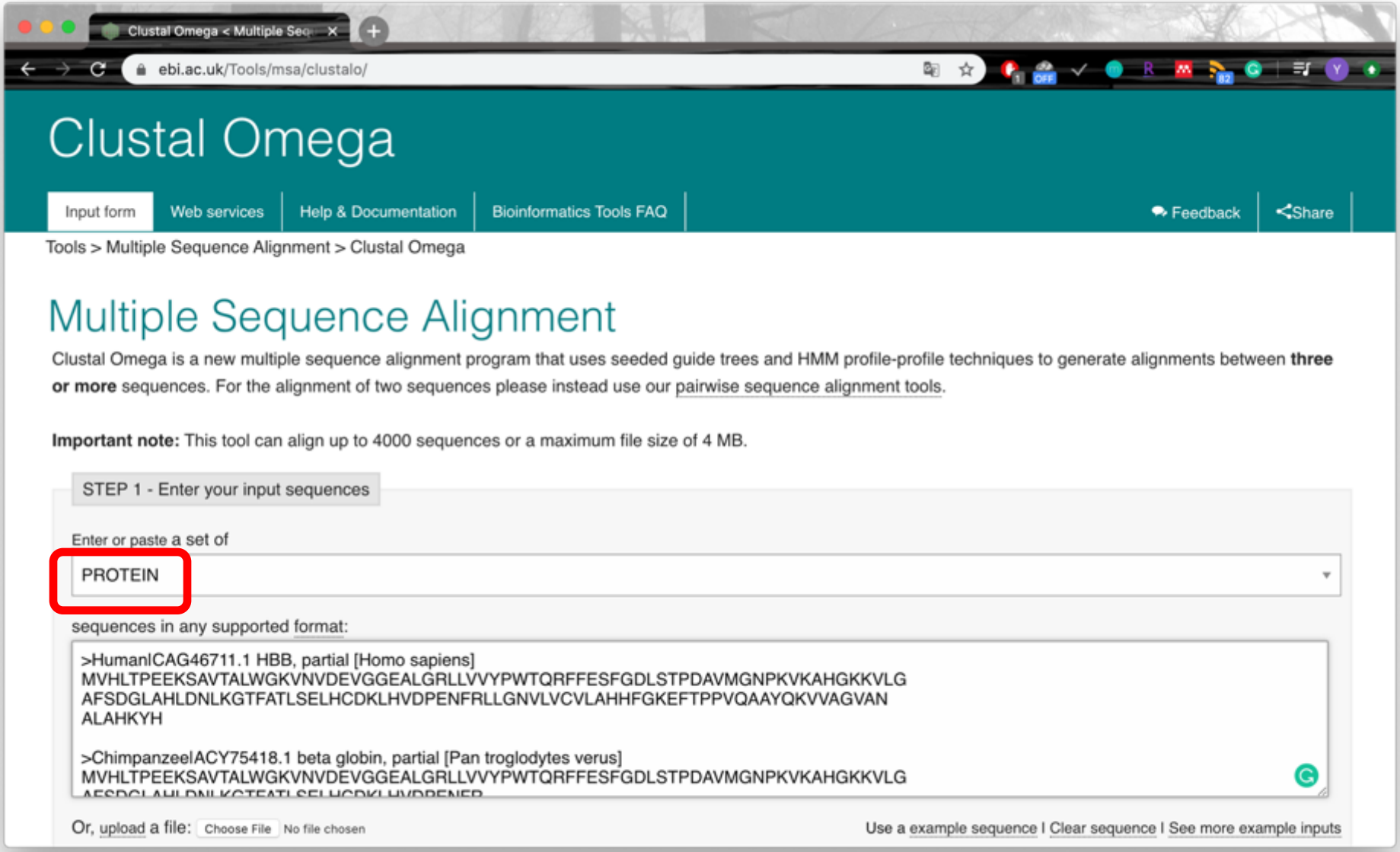
>**Mouse|**BAG16710.1 hemoglobin beta chain subunit [Mus musculus]
MVHLTDAEKAAVSGLWGKVNSDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAKVKAHGKKVIT
AFNEGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLGKDFTPAAQAAFQKVMAGVAT
ALAHKYH

>**Mouse|**AJQ22215.1 beta-globin [Peromyscus maniculatus]
MVHLTDAEKALVTGLWGKVKPEEIGGEALGRLLAVYPWTQRFFDSFGDLSSASAIMGNAKVKGHGKKVID
SFGEGLKHLDNLKGTFASLSELHCDKLHVDPENFKLLGNMIVIVMAHHLGKDFTPAAQAAYQKVVAGVAT
ALAHKYH

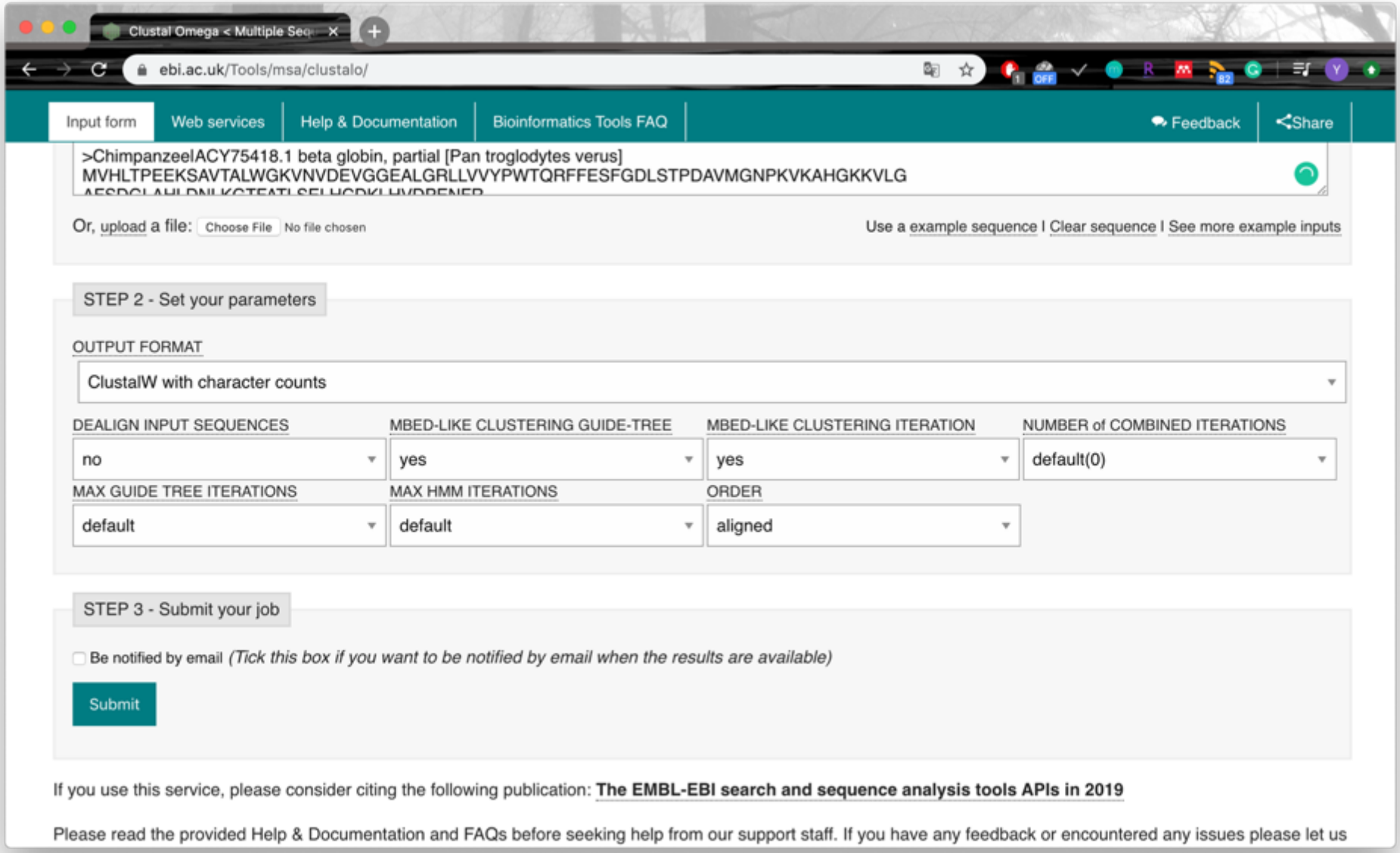
Find the tool for MSA: Clustal Omega



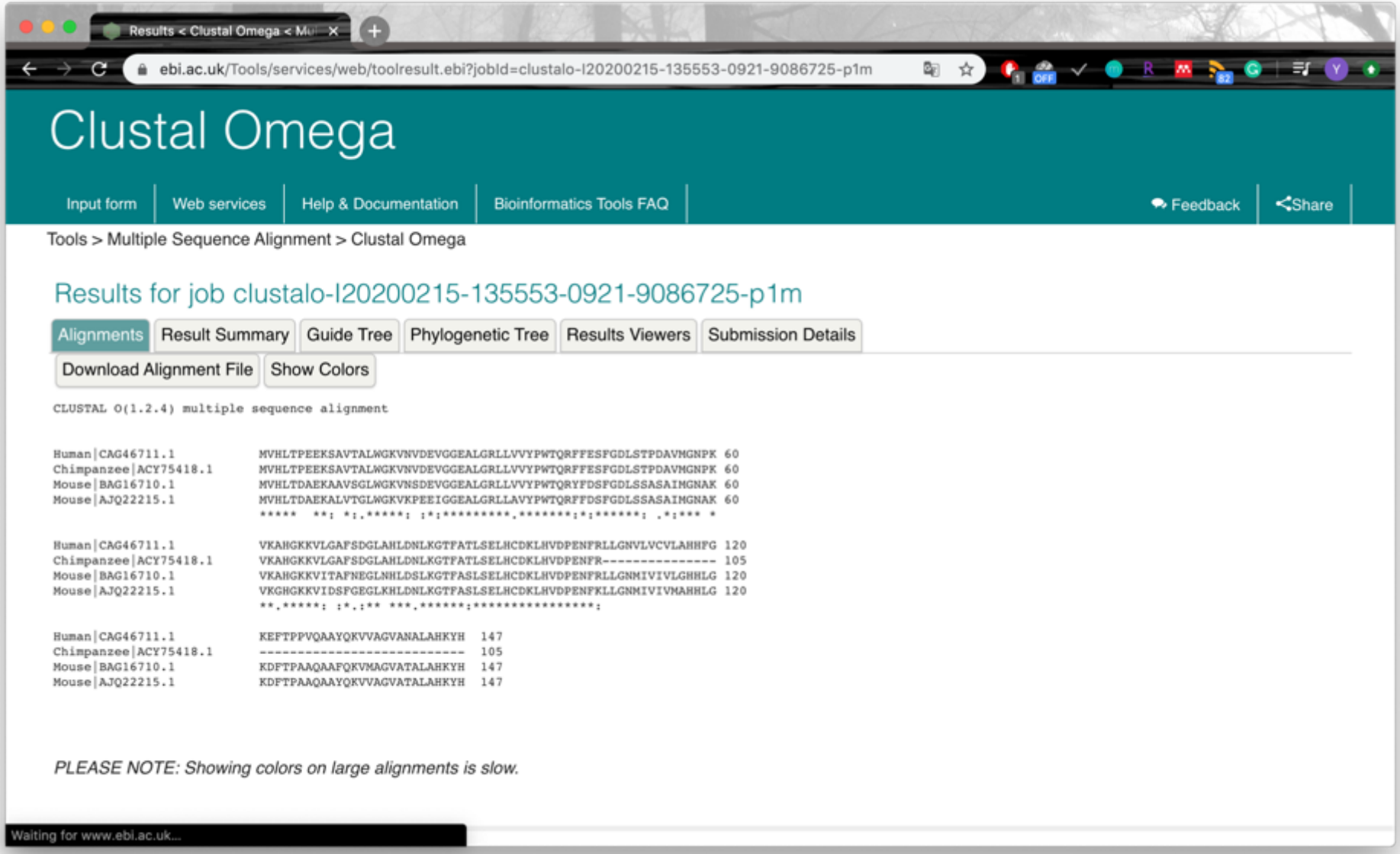
Input the sequences



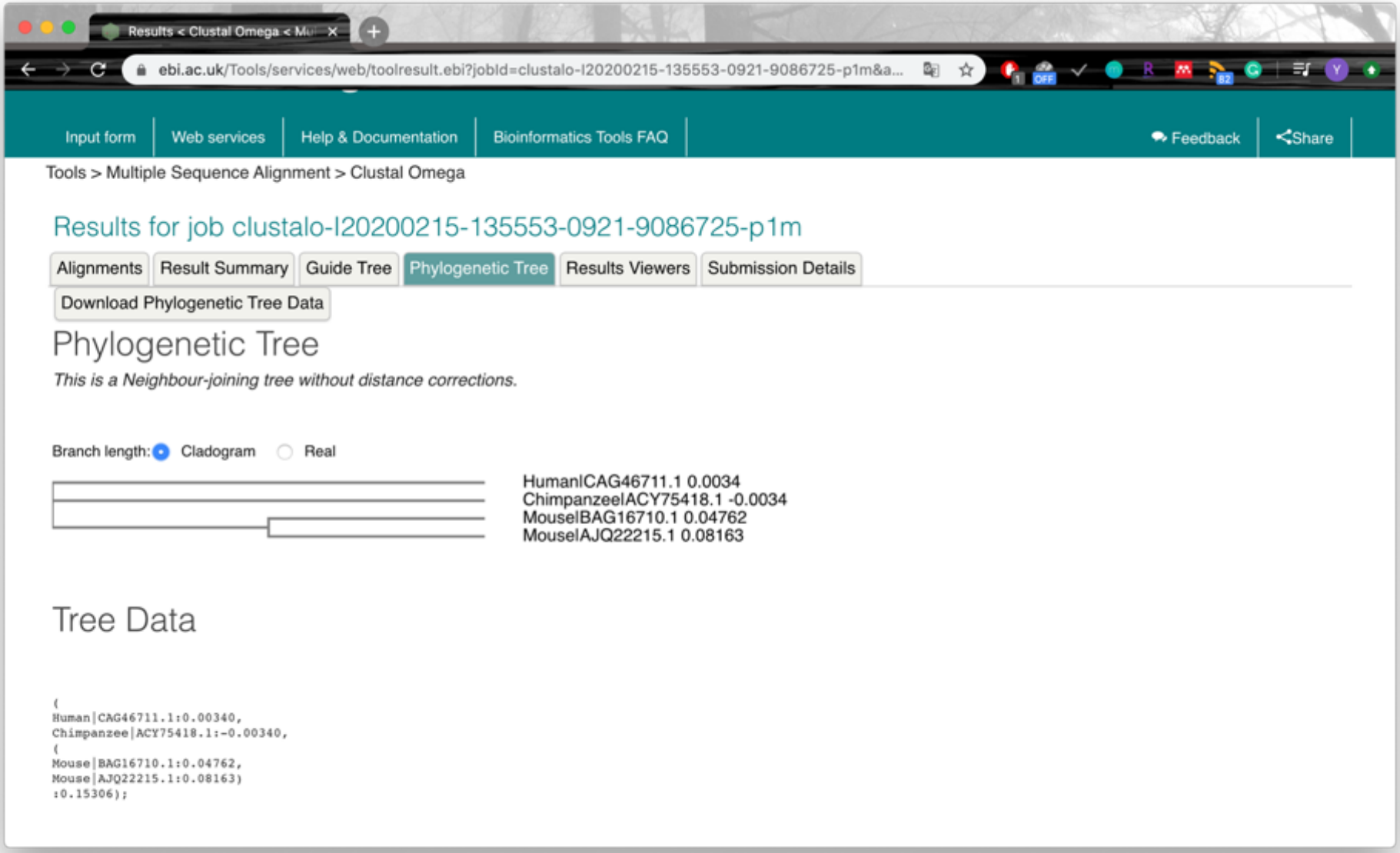
Use the default parameters



Show the alignment



Show the tree



Check list

- What are the differences between dynamic programming and heuristics searching in solving the sequence alignment problem?
- What are the steps in performing FASTA, BLAST and MSA, respectively?
- What is E-value?
- What are the differences among different versions of BLAST?