



Introduction to bioinformatics

lab1 - Alignment and Phylogenetic tree

Jiaxing Chen

The National Center for Biotechnology Information (NCBI), also known as the National Library of Medicine, is one of the most influential bioinformatics databases globally. It develops various types of databases and provides online tools for bioinformatics analysis. All NCBI's databases and software programs can be downloaded from its anonymous FTP server.

National Library of Medicine
National Center for Biotechnology Information

All Databases

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases


Download
Transfer NCBI data to your computer


Learn
Find help documents, attend a class or watch a tutorial


Develop
Use NCBI APIs and code libraries to build applications


Analyze
Identify an NCBI tool for your data analysis task


Research
Explore NCBI research and collaborative projects


Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Which animals can catch and transmit human viral infections? 29 Aug 2023

Using the NIH Comparative Genomics Resource (CGP) to understand 29 Aug 2023

Improvements to the Genetic Testing Registry (GTR®) Submission Portal 24 Aug 2023

Thank you for your feedback! You asked, we listened! In response to your 24 Aug 2023

NCBI Hidden Markov Models (HMM) Release 13.0 Now Available! 22 Aug 2023

Release 13.0 of the NCBI protein profile Hidden Markov models (HMMs) used by 22 Aug 2023

[More...](#)

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

1. Sequence query

BLAST (Basic Local Alignment Search Tool)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Step 1. Access the NCBI BLAST homepage;

Step 2. Select the corresponding BLAST program based on the input content, selecting BLASTn here.

The screenshot shows the NCBI BLAST homepage. At the top, there's a blue header bar with the NIH logo, the text "National Library of Medicine", and "National Center for Biotechnology Information". Below the header, the word "BLAST®" is displayed. In the top right corner, there are links for "Home", "Recent Results", "Saved Strategies", and "Help". A user profile icon is also present. On the left side, under the heading "Basic Local Alignment Search Tool", there's a brief description of what BLAST does: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." Below this description is a link "Learn more". To the right of the description, there's a green vertical bar labeled "NEWS" with the text "Try BLAST+ 2.14.1 today! Check out the changes we made. Tue, 22 Aug 2023" and a link "More BLAST news...". Below the main description, there are three main options: "Nucleotide BLAST" (highlighted with a red border), "blastx" (translated nucleotide ➤ protein), and "tblastn" (protein ➤ translated nucleotide). To the right of these is "Protein BLAST" (protein ➤ protein). At the bottom, there's a section for "BLAST Genomes" with a search bar and dropdown menu showing "Human", "Mouse", "Rat", and "Microbes".

Step 3. Enter the nucleic acid sequence in the QUERY section.

The screenshot shows the NCBI BLASTn suite interface. At the top, the NIH National Library of Medicine logo and the text "National Center for Biotechnology Information" are visible. Below the logo, the title "BLAST® > blastn suite" is displayed, along with links for "Home", "Recent Results", "Saved Strategies", and "Help". The main search area is titled "Standard Nucleotide BLAST". The "blastn" tab is selected. In the "Enter Query Sequence" section, a sequence of DNA is entered into the text input field, which is highlighted with a red border. This sequence is: GTGCTGCCGTACTACACAAACTGCTTGCAGTGACAATGCGTTAGCTTA CTACAAACAAACAAAGGGAGTAGTTGACTTGCACTGTATCCGATTAC AGGATTGAAATGGGCTAGATTCCCTAAGAGTGATGGAACCTGGTACTATCAT ACAGAACTGGAACCACCTTGAGTTGTTACAGACACCTAAAGGTCTAA. Below the query sequence, there are fields for "Query subrange" with "From" and "To" inputs, and a "Job Title" input field. A checkbox for "Align two or more sequences" is present. The "Choose Search Set" section includes a "Database" dropdown set to "Standard databases (nr etc.)", a "Try experimental taxonomic nt databases" link, and a "Nucleotide collection (nr/nt)" dropdown. Other search parameters like "Organism", "Exclude", "Limit to", and "Entrez Query" are also shown.

Figure show searching of COVID-19 *Orf1ab* gene as an example
Gene detail in <https://www.ncbi.nlm.nih.gov/gene/43740578>

Please try searching with sequences in the file “Manual_Nucl.fasta” from ispace

Step 4. Select search set and corresponding parameters;

Choose Search Set

Database

Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Now Experimental databases

Try experimental taxonomic nt databases [Download](#)

For more info see [What are taxonomic nt databases?](#)

Nucleotide collection (nr/nt) [?](#)

Organism **Optional**

Enter organism name or id—completions will be suggested exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude **Optional**

Model (XM/XP) Uncultured/environmental sample sequences

Sequences from type material

Limit to **Optional**

Entrez Query **Optional**

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for

Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST

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

Note: Parameter values that differ from the default are highlighted in yellow and marked with * sign

+ Algorithm parameters

Step 5. Click “BLAST” to submit the BLASTn query

Choose Search Set

Database

Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus
New Experimental databases

Try experimental taxonomic nt databases [Download](#)

For more info see [What are taxonomic nt databases?](#)

Nucleotide collection (nr/nt) [?](#)

Organism Optional

Enter organism name or id—completions will be suggested exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional

Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional

Sequences from type material

[YouTube](#) [Create custom database](#)

Entrez Query Optional

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for

Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST

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

Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign

+ Algorithm parameters

[Edit Search](#)

[Save Search](#)

[Search Summary](#)

[How to read this report?](#)

[BLAST Help Videos](#)

[Back to Traditional Results Page](#)

Job Title **Nucleotide Sequence**

RID **FBFD7Y2K016** Search expires on 09-05 19 21 pm [Download All](#)

Program **BLASTN** [?](#) [Citation](#)

Database nt [See details](#)

Query ID lcl|Query_44933

Description None

Molecule type dna

Query Length 490

Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

[Filter](#)

[Reset](#)

[Descriptions](#)

[Graphic Summary](#)

[Alignments](#)

[Taxonomy](#)

Sequences producing significant alignments

[Download](#)

[Select columns](#)

Show 100

[?](#)

select all 100 sequences selected

[GenBank](#)

[Graphics](#)

[Distance tree of results](#)

[MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/OKPHL0027690/2023 ORF1ab	Severe acute res...	884	884	100%	0.0	100.00%	29822	OR480076.1
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 genome assembly_complete genome_monopartite	Severe acute res...	884	884	100%	0.0	100.00%	29790	OY639170.1
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 genome assembly_complete genome_monopartite	Severe acute res...	884	884	100%	0.0	100.00%	29786	OY639141.1
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 genome assembly_complete genome_monopartite	Severe acute res...	884	884	100%	0.0	100.00%	29786	OY639123.1
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 genome assembly_complete genome_monopartite	Severe acute res...	884	884	100%	0.0	100.00%	29786	OY639116.1
<input checked="" type="checkbox"/>	Severe acute respiratory syndrome coronavirus 2 genome assembly_complete genome_monopartite	Severe acute res...	884	884	100%	0.0	100.00%	29786	OY639070.1

From this page, we can obtain the following main biological information:

RID: Unique identifier assigned to this search.

DESCRIPTIONS: Default display options for BLAST results, where the description of each search result serves as a brief name of the search object.

SCIENTIFIC NAME: Species information corresponding to the search results.

QUERY COVER: The coverage between the target sequence and the query sequence.

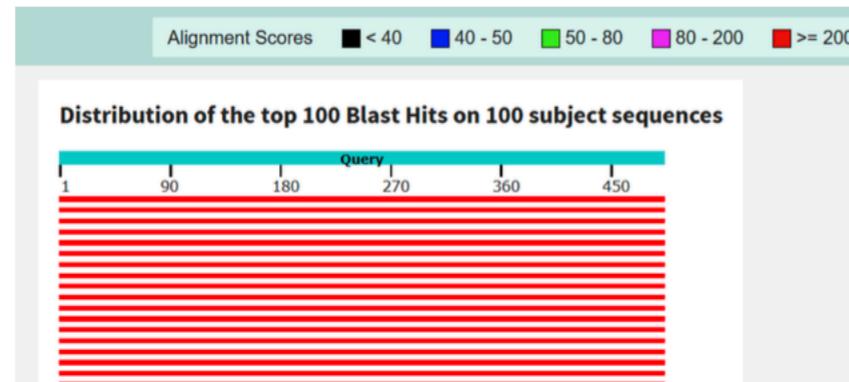
E-VALUE: The reliability of the alignment of the reaction sequence, with smaller E values indicating higher reliability.

PER.IDENT: The value of sequence alignment consistency. (Percentage identity)

ACC.LEN: Target sequence length.

ACCESSION: Accession ID of target sequence in NCBI database.

GRAPHIC SUMMARY: Visually presenting the alignment results with similar sequences.



ALIGNMENTS: The results/status of sequence alignment, including consistency, similarity, sequence length, etc.

Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/OKPHL0027690/2023 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), ORF6 protein (ORF6), ORF7a protein (ORF7a), and ORF7b (ORF7b) genes, complete cds; ORF8 gene, complete sequence; and nucleocapsid phosphoprotein (N) and ORF10 protein (ORF10) genes, complete cds

Sequence ID: [OR480076.1](#) Length: 29822 Number of Matches: 1

Range 1: 12701 to 13190 GenBank Graphics					▼ Next Match	▲ Previous Match
Score 884 bits(980)	Expect 0.0	Identities 490/490(100%)	Gaps 0/490(0%)	Strand Plus/Plus		
Query 1	GTGCTGCCGGTACTACACAAACTGCTTCACTGATGACAATGCGTTAGCTTACTACAACA		60			
Sbjct 12701	GTGCTGCCGGTACTACACAAACTGCTTCACTGATGACAATGCGTTAGCTTACTACAACA		12760			
Query 61	CAACAAAGGGAGGTAGGTTTGACTTGCACGTGTTATCCGATTTACAGGATTGAAATGGG		120			
Sbjct 12761	CAACAAAGGGAGGTAGGTTTGACTTGCACGTGTTATCCGATTTACAGGATTGAAATGGG		12820			
Query 121	CTAGATTCCCTAACAGAGTGAATGGAACGTGGTACTATCTATACAGAACCTGGAAACCACCTTGT		180			
Sbjct 12821	CTAGATTCCCTAACAGAGTGAATGGAACGTGGTACTATCTATACAGAACCTGGAAACCACCTTGT		12880			
Query 181	GGTTTGTTACAGACACACCTAAAGGTCTAAAGTGAAGTATTATACTTTATTAAAGGAT		240			
Sbjct 12881	GGTTTGTTACAGACACACCTAAAGGTCTAAAGTGAAGTATTATACTTTATTAAAGGAT		12940			
Query 241	TAAACAAACCTAAATAGAGGTATGGTACCTGGTAGTTAGCTGCCACAGTACGTCACAAAG		300			
Sbjct 12941	TAAACAAACCTAAATAGAGGTATGGTACCTGGTAGTTAGCTGCCACAGTACGTCACAAAG		13000			

TAXONOMY: Classification of similar sequences within species.

100 sequences selected ?				
Organism	Blast Name	Score	Number of Hits	Description
Severe acute respiratory syndrome coronavirus 2	viruses	884	101	Severe acute respiratory syndrome co

2. Pairwise Sequence Alignment

METHOD 1: Use BLAST pairwise alignment

URL: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Step 1. Select BLAST mode (Please refer to **Lab1 - Introduction to NCBI Database and BLAST for the content**. Taking BLASTn as an example)

The screenshot shows the NCBI BLAST homepage. At the top, there's a banner for the ClusteredNR database on BLAST+. Below it, the "Basic Local Alignment Search Tool" is highlighted. A news box on the right announces the availability of the ClusteredNR database on BLAST+. The "Web BLAST" section features three main search tools: Nucleotide BLAST (nucleotide to nucleotide), blastx (translated nucleotide to protein), and tblastn (protein to translated nucleotide). The Protein BLAST tool is also shown. At the bottom, there's a "BLAST Genomes" search bar with options for Human, Mouse, Rat, and Microbes.

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

Check out the ClusteredNR database on BLAST+ [Learn more](#) [Give us feedback](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS ClusteredNR database on BLAST+
The ClusteredNR database is now available for BLAST+
Thu, 24 Aug 2023 [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Human Mouse Rat Microbes [Search](#)

Step 2. Click “Align two or more sequences”

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

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

Query subrange [?](#)

From
To

Or, upload file 未选择文件. [?](#)

Job Title

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

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

New Experimental databases [Try experimental taxonomic nt databases](#) [Download](#)

For more info see [What are taxonomic nt databases?](#)

Nucleotide collection (nr/nt) [?](#)

Organism Optional

Enter organism name or id—completions will be suggested exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional

Models (XM/XP) Uncultured/environmental sample sequences

Sequences from type material

Limit to Optional

Entrez Query Optional

Enter an Entrez query to limit search [?](#) [YouTube](#) Create custom database

Step 3. Input query sequence (or NCBI ID) & target sequence (or NCBI ID) (Taking 2 parts of hAPOBEC3A coding sequence, NCBIID: NM_001270406.2 as example)

Align Sequences Nucleotide BLAS I

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. mor

Enter Query Sequence

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

```
TTTCAATGTATAAT  
GAAATGAAATGATAATTGGCTTCATATCTAGACTAACACAAAATTAGAACATCT  
TCCATAATTGCTTTG  
CTCAGTAACGTGTCAATTGCAAGAGTTCCACAAACACTA
```

From
To

Or, upload file 未选择文件.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

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

```
TACTAAAAACAAAAATTAGCCAGGCCTGGTGGCGGGCGCCTGTAGTCC  
CAGCTACTCTGGAGGCTGA  
GGCAGGAGAGTAGCGTGAACCCGGGAGGCAGAGCTGCGGTGAGCCGA  
GATTGCGCTACTGCACTCCAGC
```

From
To

Or, upload file 未选择文件.

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Show results in a new window

Step 4. Click “BLAST”

Align Sequences Nucleotide BLAST

blastn **blastp** **blastx** **tblastn** **tblastx**

BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)
TTTCAATGTAATTAAI
GAAATGAAATGATAATTGGCTTCATATCTAGACTAACACAAAATTAAGAATCT
TCCATAATTGCTTTG
CTCAGTAACTGTGTCAATTGCAAGAGTTCCACAAACACTA

Query subrange [?](#)
From
To

Or, upload file [浏览...](#) 未选择文件. [?](#)

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

Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)
TACTAAAAATACAAAAAAATTAGCCAGGCGTGGCGGGCGCCTGTAGTCC
CAGCTACTCTGGAGGCTGA
GGCAGGAGAGTAGCGTGAACCCGGGAGGCAGAGCTTGCGGTGAGCCGA
GATTGCGCTACTGCACTCCAGC

Subject subrange [?](#)
From
To

Or, upload file [浏览...](#) 未选择文件. [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST Search nucleotide sequence using Megablast (Optimize for highly similar sequences) Show results in a new window

+ Algorithm parameters

Step 5. Result

◀ Edit Search Save Search Search Summary ▾

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title Nucleotide Sequence

RID H3BKW59N114 Search expires on 09-26 23:59 pm
[Download All](#) ▾

Program Blast 2 sequences [Citation](#) ▾

Query ID Icl|Query_74263 (dna)

Query Descr None

Query Length 394

Subject ID Icl|Query_74265 (dna)

Subject Descr None

Subject 210

Length

Other reports [MSA viewer](#) ?

Filter Results

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Dot Plot

Sequences producing significant alignments Download Select columns Show 100 ?

select all 1 sequences selected [Graphics](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	None provided		388	388	53%	6e-113	100.00%	210	Query_74265

METHOD 2: Use EMBOSS pairwise alignment toolbox

URL: <https://www.ebi.ac.uk/Tools/psa/>

Step 1. Select algorithm (Here, take Needleman-Wunsch as example)

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, **Multiple Sequence Alignment (MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned.

Needle (EMBOSS)

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

[Launch !\[\]\(8706f9f9febc74216a91030d11f10ce7_img.jpg\)Needle](#)

Stretcher (EMBOSS)

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

[Launch !\[\]\(b65ff707ec4d1ab514bcb3ba54feee42_img.jpg\)Stretcher](#)

GGSEARCH2SEQ

GGSEARCH2SEQ finds an optimal global alignment using the Needleman-Wunsch algorithm.

[Launch !\[\]\(9321542a4f5100bf6d54b71e7f20e8cc_img.jpg\)ggsearch2seq](#)

Step 2. Switch align mode (Enter a pair of: XX) to “DNA”

Enter a pair of

PROTEIN
PROTEIN
DNA

Or, upload a file: 未选择文件。

Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **protein** sequence in any supported format:

Or, upload a file: 未选择文件。

STEP 2 - Set your pairwise alignment options

Step 3. Input query sequence & target sequence (Taking hAPOBEC3A coding sequence, NCBIID: KM266646.1 & rAPOBEC3A coding sequence NCBIID: NM_001033703.2 as example)

(Search in NCBI to get fasta for KM266646.1 and NM_001033703.2

<https://www.ncbi.nlm.nih.gov/nuccore/KM266646.1?report=fasta>

https://www.ncbi.nlm.nih.gov/nuccore/NM_001033703.2?report=fasta

Pairwise Sequence Alignment

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

STEP 1 - Enter your protein sequences

Enter a pair of

DNA

sequences. Enter or paste your first **protein** sequence in any supported format:

```
>NM_001033703.2 Rattus norvegicus apolipoprotein B mRNA editing enzyme catalytic subunit 3 (Apobec3), mRNA
GGAAGTCACTCGAACTTCTGGGGTCTCCCAAAGGCCAGGGCCTGTACATTGGCAGTTGTACAAATGCAACC
CCAGGGTCTGGGGCCAACGCTGGGATGGGACCATGTGCCTGGGATGCAGCCATGCAGACCCCTATTCA
CCGATCAGAAACCCGCTAAAGAACGTTATATCACAAACATTCTACTTTCTATTAAAGAACGTTACGCTATG
CCTGGGGTCGAAAGATAACTTCTTGCTATGAAGTGAATGGGATGGACTGCCTTACCTGTCCCCCT
TCGCCAAGGGGTCTTCAGGAAACAGGGCACATCCACGCCGAACCTGCTTCATATACTGGTCCACGAC
AAAGTCTGAGAGTGCTGTCCCCGATGGAAGAGTCAAGGTCACGTGGTACATGTCTGGAGCCCCCTGCA
GCAACTCCCCCAGCCACCTACCCACCTCCCCACACCCCCAACCTAACCCCCATCTCACCTC
```

Or, upload a file: 未选择文件。

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

AND

Enter or paste your second **protein** sequence in any supported format:

```
>KM266646.1 Homo sapiens APOBEC3A (APOBEC3A) mRNA, complete cds
ATGGAAGCCAGCCCAGCATCCGGGCCAGACACTTGATGGATCCACACATATTCACTTCAACTTTAAC
ATGGCATTGGAAGGCATAAGACCTACCTGTGCTACGAAGTGGAGCGCCTGGACAATGGCACCTCGGTCAA
GATGGACCAGCACAGGGGTTTCTACACAACCAGGCTAAGAACATTTCTGTGGCTTTACGGCCGCCAT
GCGGAGCTGCGCTTCTGGACCTGGTTCTTGCAGTGGACCCGGCCAGATCTACAGGGTCACCT
GGTCATCTCTGGAGCCCCCTGCTTCTCTGGGGCTGTGCCGGGGAAAGTGCCTGCGTTCTCAGGAGAA
CACACACGTGAGACTGCGCATCTCGCTGCCGCATCTATGATTACGACCCCCCTATATAAGGGAGGGCTG
GAAATCTCCCCCTACCTCTCCCCCTACCTCTCCCCCTACCTACCTACCTACCTCTCCCCCTACCTCT
```

Or, upload a file: 未选择文件。

Step 3. Click “SUBMIT”

STEP 2 - Set your pairwise alignment options

OUTPUT FORMAT

pair

The default settings will fulfill the needs of most users.

More options... (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

If you use these services, please consider citing the following publication: [Search and sequence analysis tools services from EMBL-EBI in 2022](#)

Step 4. Results

Results for job emboss_needle-I20230925-170034-0785-96430502-p1m

Alignment Submission Details

[View Alignment File](#)

```
#####
# Program: needle
# Rundate: Mon 25 Sep 2023 17:00:36
# Commandline: needle
#   -auto
#   -stdout
#   -asequence emboss_needle-I20230925-170034-0785-96430502-p1m.asequence
#   -bsequence emboss_needle-I20230925-170034-0785-96430502-p1m.bsequence
#   -datafile EDNAFULL
#   -gapopen 10.0
#   -gapextend 0.5
#   -endopen 10.0
#   -endextend 0.5
#   -aformat3 pair
#   -snucleotide1
#   -snucleotide2
# Align_format: pair
# Report_file: stdout
#####

=====
#
# Aligned_sequences: 2
# 1: NM_001033703.2
# 2: KM266646.1
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1855
# Identity: 400/1855 (21.6%)
# Similarity: 400/1855 (21.6%)
# Gaps: 1354/1855 (73.0%)
# Score: 857.0
#
#
=====
```

NM_001033703.	1	GGAAGTCACTCGAACCTCTGGGTCTCCCAAAGCCAGGGCTGTACATTG	50
KM266646.1	1	-----	0
NM_001033703.	51	GCAGTTGTACAAATGCAACCCCAGGGTCTGGGCCAACGCTGGGATGGG	100
		.	
KM266646.1	1	-----ATGGA	5
NM_001033703.	101	A-CCAGTGTGCCTGGGATGCAGCCATCGCAGACCCATTACCGATCAGA	149
		.	

3. Align multiple sequence

TOOL: EMBOSS MSA toolbox

<https://www.ebi.ac.uk/jdispatcher/msa>

Step 1. Select algorithm (Here, take Clustal-Omega as example)

Tools > Multiple Sequence Alignment

Service Announcement

The new Job Dispatcher Services beta website is now available at <https://wwwdev.ebi.ac.uk/Tools/jdispatcher>. We'd love to hear your feedback about the new webpages!

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

 [Launch Clustal Omega](#)

Cons (EMBOSS)

EMBOSS Cons creates a consensus sequence from a protein or nucleotide multiple alignment.

 [Launch EMBOSS Cons](#)

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

 [Launch Kalign](#)

Step 2. Switch align mode (Enter a pair of: XX) to “DNA”

Enter or paste a set of

PROTEIN
PROTEIN
DNA
RNA

Step 3. Paste multiple sequence or upload files

STEP 1 - Enter your input sequences

Enter or paste a set of

DNA

sequences in any supported format:

```
>KM266646.1 Homo sapiens APOBEC3A (APOBEC3A) mRNA, complete cds
ATGGAAGCCAGCCCAGCATTGGATCCACACATATTCACTTCAAACTTAAC
ATGGCATTGGAAGGCATAAGACCTACCTGTGCTACGAAGTGGAGCGCCTGGACAATGGCACCTCGGTCAA
GATGGACCAGCACAGGGCTTCTACACAACCAGGCTAAGAATCTTCTGTGGCTTACGGCCGCCAT
GCGGAGCTGCGCTTGGACCTGGTCTTGCAGTTGGACCCGGCCAGATCTACAGGGTCACTT
GGTCATCTCCTGGAGCCCCCTGCTTCTCCTGGGGCTGTGCCGGGGAAAGTGCCTGCCTTCAGGAGAA
CACACACGTGAGACTGCGCATCTCGCTGCCCGCATCTATGATTACGACCCCCCTATATAAGGAGGCGCTG
CAAACTCTCCCCCATCTCCCCCAACTCTCCATCATCACCTACCATCTTACCTCTCCCCACAA
```

Or, upload a file: 未选择文件。

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

Step 4. Click “SUBMIT”

STEP 3 - Submit your job

Be notified by email (*Tick this box if you want to be notified by email when the results are available*)

Submit

Step 5. Result

Results for job clustalo-l20230926-045902-0090-72510752-p1m

Alignments **Result Summary** **Guide Tree** **Phylogenetic Tree** **Results Viewers** **Submission Details**

[Download Alignment File](#)

CLUSTAL O(1.2.4) multiple sequence alignment

XM_039462978.1	-	0
HQ404235.1	-	0
HQ404234.1	-	0
HQ404233.1	-	0
KM266650.1	-	0
JF831054.1	-	0
KX583650.1	-	0
XM_058431894.1	-	0
XM_024239990.2	-	0
XM_054470453.1	-	0
XM_030815195.1	-	0
KM266646.1	-	0
XM_034949099.3	CCGATGATATATTAAGGCTCCT-	22
NM_001163936.1	AAAAAGGAGGAGGACACAGGCTGTGACTGAGCATACCATCAGAGGACTCAGAGGCCAGGCC	60
XM_007189074.3	-	0
XM_036455790.2	-	0
XM_042781063.1	-	0
XM_039462978.1	-	0
HQ404235.1	-	0
HQ404234.1	-	0
HQ404233.1	-	0
KM266650.1	-	0
JF831054.1	-	0
KX583650.1	-	0
XM_058431894.1	-	0
XM_024239990.2	-	0
XM_054470453.1	-	0
XM_030815195.1	-	0
KM266646.1	-	0
XM_034949099.3	-GAATCCTAAG-----AGAATTTGGTGAAGATCTAACACCAC	60
NM_001163936.1	TGCCGCTTGAACAACCTCAAGGAGGAGGCCACAGGCTGTGACTGAGCATACCATCAGAGG	120
XM_007189074.3	GCTGTGACCAAGAACAGAAACTGAAG	26
XM_036455790.2	-	0
XM_042781063.1	-	0

4. Construct Phylogenetic tree

Here, the phylogenetic tree (& nwk text) generated from the MSA will be directly presented on the EMBOSS result page.

Results for job clustalo-l20230926-045902-0090-72510752-p1m

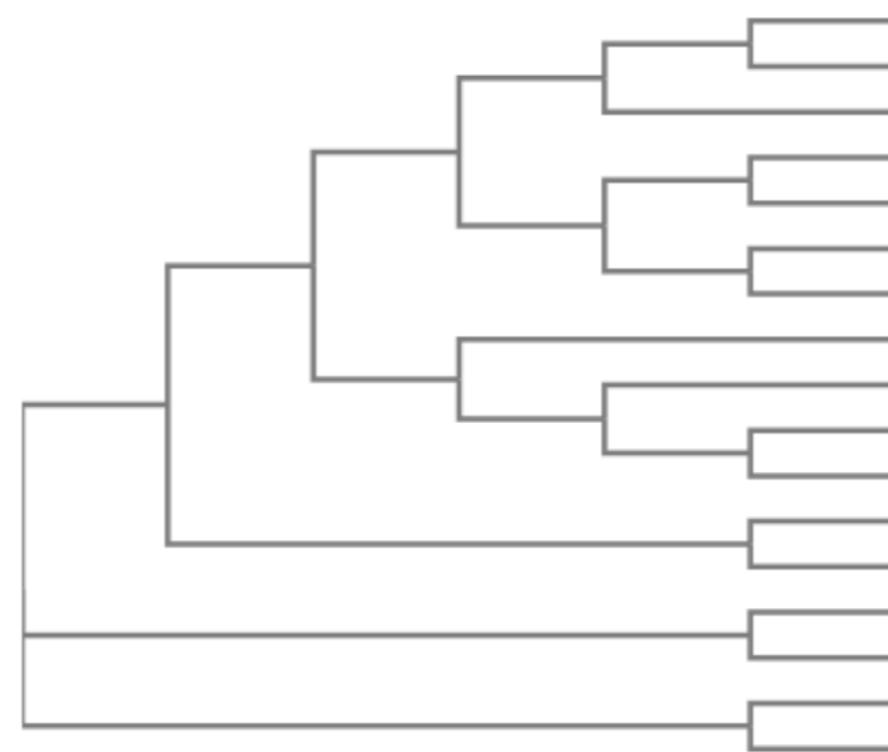
Alignments Result Summary Guide Tree **Phylogenetic Tree** Results Viewers Submission Details

[Download Alignment File](#)

CLUSTAL O(1.2.4) multiple sequence alignment

XM_039462978.1	-----	0
HQ404235.1	-----	0
HQ404234.1	-----	0
HQ404233.1	-----	0
KM266650.1	-----	0
JF831054.1	-----	0
KX583650.1	-----	0
XM_058431894.1	-----	0
XM_024239990.2	-----	0
XM_054470453.1	-----	0
XM_030815195.1	-----	0
KM266646.1	-----	0
XM_034949099.3	-CCGATGATATATTAAAGGCTCCT-	22
NM_001163936.1	AAAAAGGAGGAGGACACAGGCTGTGACTGAGCATACCATCAGAGGACTCAGAGGCCAGGCC	60
XM_007189074.3	-----	0
XM_036455790.2	-----	0
XM_042781063.1	-----	0
XM_039462978.1	-----	0
HQ404235.1	-----	0
HQ404234.1	-----	0
HQ404233.1	-----	0
KM266650.1	-----	0
JF831054.1	-----	0
KX583650.1	-----	0
XM_058431894.1	-----	0
XM_024239990.2	-----	0
XM_054470453.1	-----	0
XM_030815195.1	-----	0
KM266646.1	-----	0
XM_034949099.3	-GAATCCTAAG-----AGAATGTTGGTGAAGATCTAACACCAC	60
NM_001163936.1	TGCCGCTTGAACAACCAAGGAGGAGGCCACAGGCTGTGACTGAGCATACCATCAGAGG	120
XM_007189074.3	-GCTGTGACCAAGAACAGAAACTGAAG-----	26
XM_036455790.2	-----	0
XM_042781063.1	-----	0

Branch length: ● Cladogram ○ Real



XM_039462978.1	0.04509
HQ404234.1	0.01018
HQ404235.1	0.02497
NM_001163936.1	0.13495
XM_007189074.3	0.08858
XM_036455790.2	0.17915
XM_042781063.1	0.13525
HQ404233.1	0.02634
KM266650.1	0.00741
JF831054.1	0.00684
KX583650.1	0.00465
KM266646.1	0.00392
XM_034949099.3	0.01275
XM_058431894.1	0.03366
XM_030815195.1	0.00539
XM_024239990.2	0.00087
XM_054470453.1	0.00206

5. Task

Background

In the above Manual module, we conducted related exercises utilizing the APOBEC3A mRNA sequences. In the Task module, we will analyze the APOBEC3A coding protein. APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) proteins have a characteristic zinc-coordination motif (H-X-E-X₂₃-28-P-C-X-C) within the active site where a water molecule binds Zn²⁺ and the metal ion is coordinated by one histidine and two cysteines. The AID/APOBEC family enzymes convert cytosines in single-stranded DNA to uracil causing base substitutions and strand breaks. Based on this property, relevant research has used human APOBEC3A (hAPOBEC3A) fusion dead Cas9 (dCas9) protein as a base editor for precise genome editing.

Task1. Using the human APOBEC3A (**hAPOBEC3A**) protein sequence (ID: **NP_663745.1**) as query, use BLASTp to search for its homologous protein sequence.

Task2. For the **APOBEC3A.fasta** file (from iSpace), perform multiple sequence alignment and construct phylogenetic tree.

Task3. Attempt to interpret the results of this phylogenetic tree.

Task4. Based on the phylogenetic tree, select the sequences closest to hAPOBEC3A and generate **pairwise sequence alignment**, attempt to interpret the specific differences between the two sequences (hAPOBEC3A and target sequence).

Closest: Pongo pygmaeus APOBEC3A protein