

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

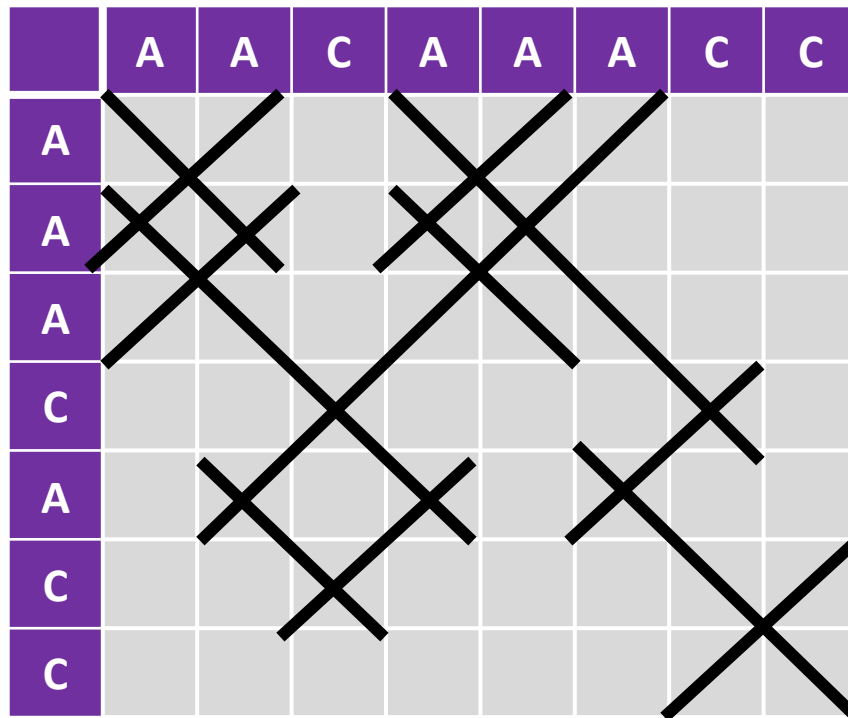
- To align two sequences of lengths m and n
 - Time complexity: $O(mn)$
 - Space complexity: $O(mn)$
- To find the sequence in a database with l length- n sequences that is closest to a query sequence of length m (Suppose we consider the database sequences one by one)
 - Time complexity: $O(lmn)$
 -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

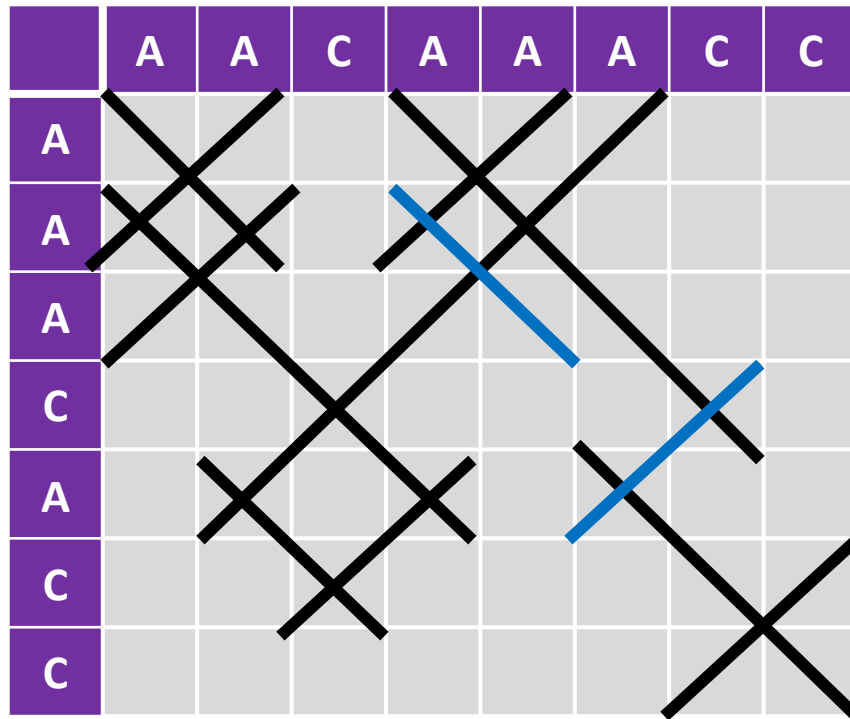
Example: Dot Plot

- A Dot Plot is provided. Answer the questions.



Example: Dot Plot

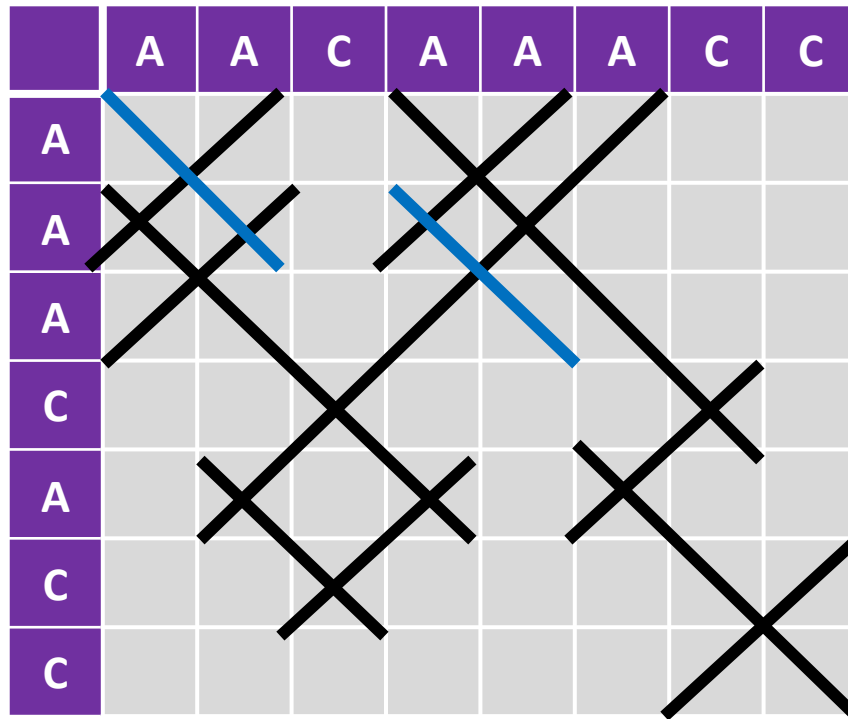
- A Dot Plot is provided. Answer the questions.



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

Example: Dot Plot

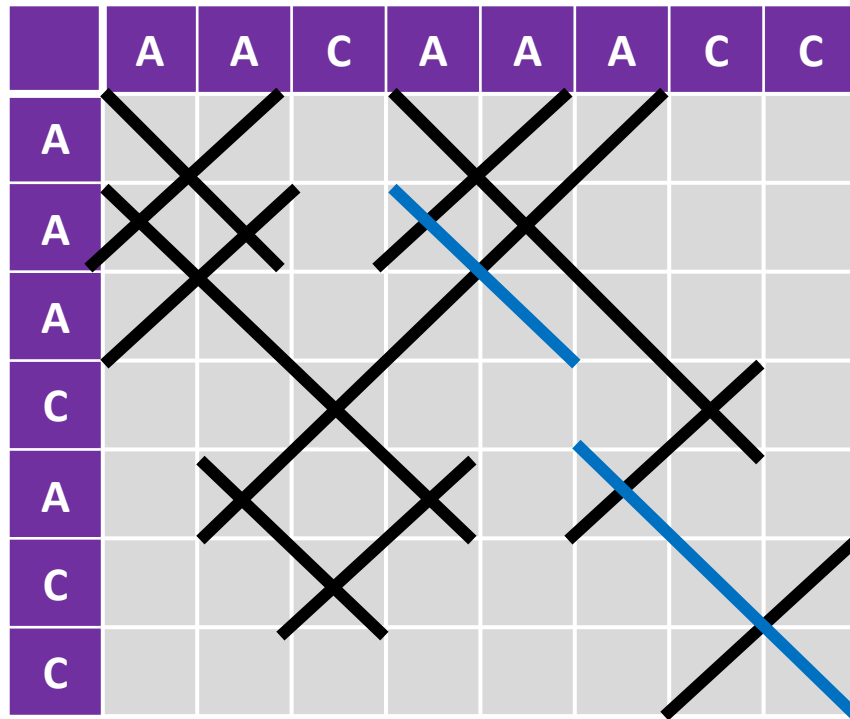
- A Dot Plot is provided. Answer the questions.



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

Example: Dot Plot

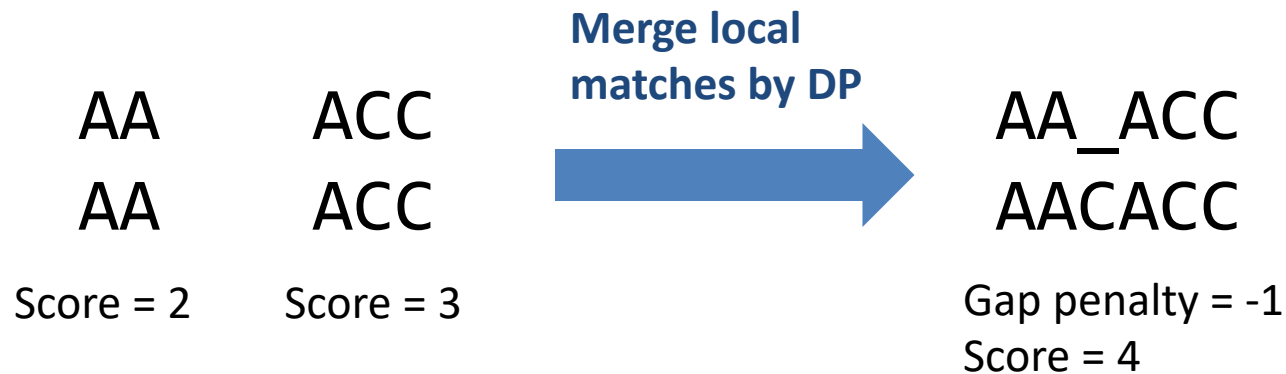
- A Dot Plot is provided. Answer the questions.



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

Example: Dot Plot

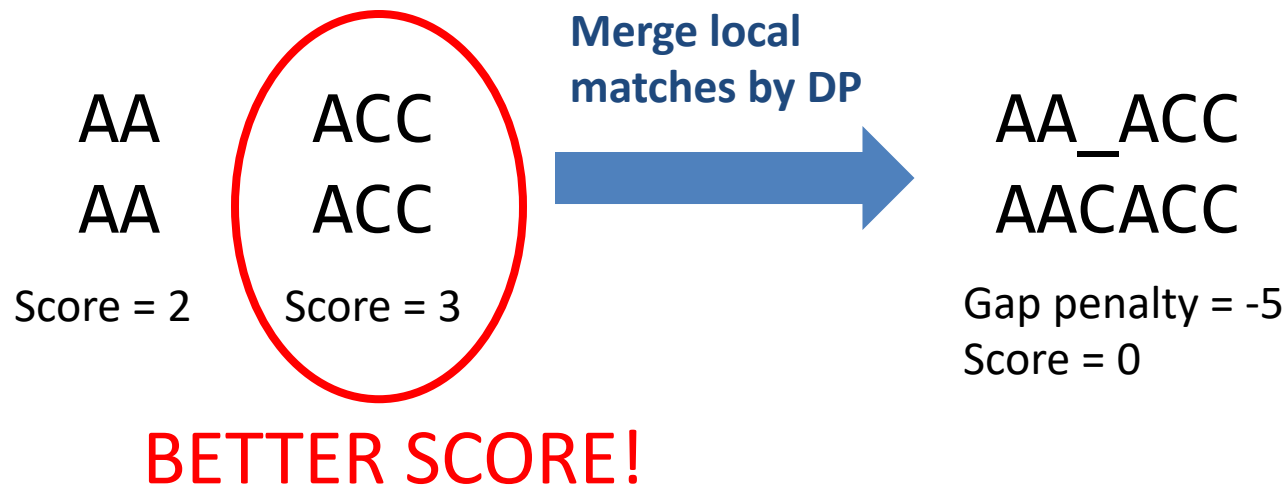
- It depends on the scoring scheme
- If merging two plots by DP will give a higher alignment score (e.g. match = 1, gap penalty = -1), we will merge the alignments



MERGE!

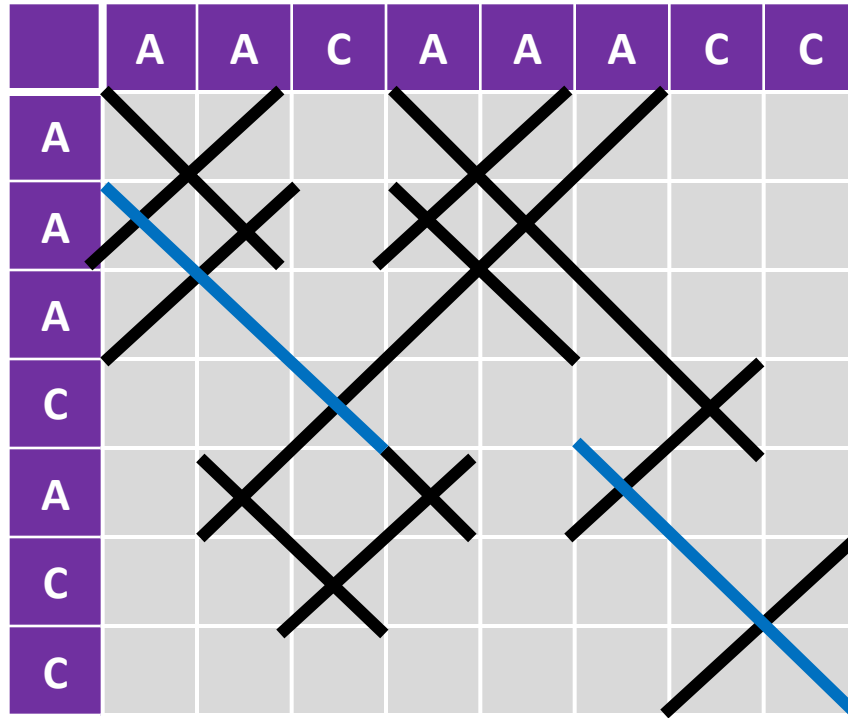
Example: Dot Plot

- 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

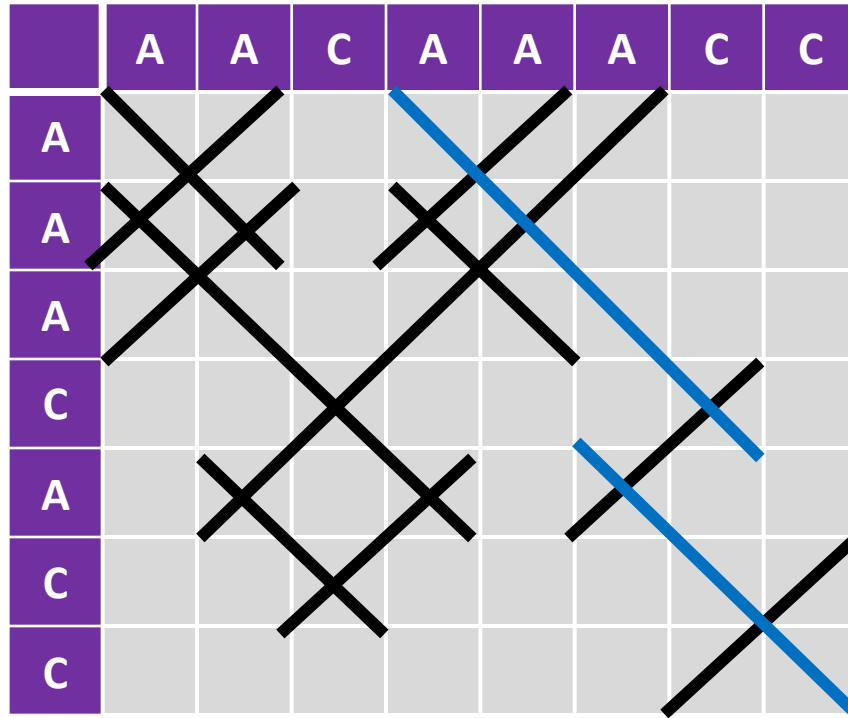
- A Dot Plot is provided. Answer the questions.



- 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

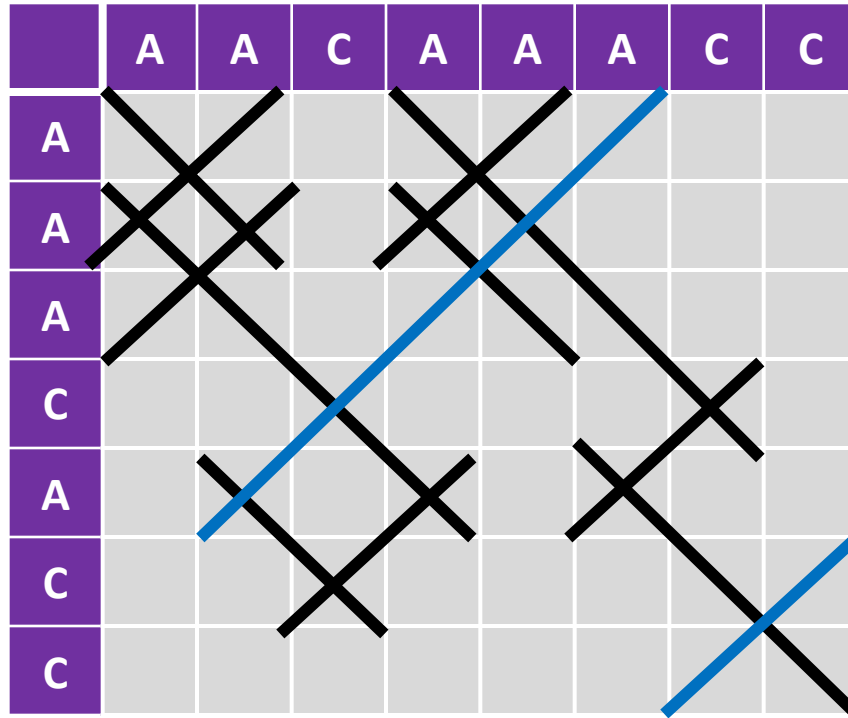
- A Dot Plot is provided. Answer the questions.



- 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
1 2 3 4 5 6 7 8 9 0 1 2 3 4
C C T A C C G A T A C C G A

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 <i>r</i>	Position
ACCGA	1
ATAGC	5
CCGAT	2
CGATA	3
GATAG	4

5-mer in <i>s</i>	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:

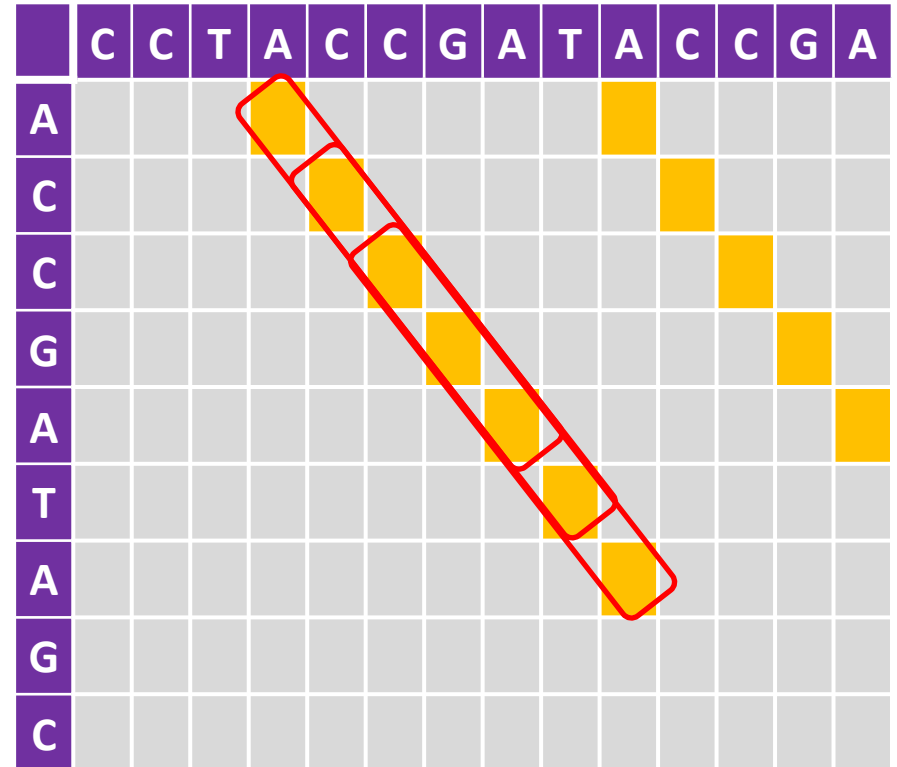
$r[1..5]=s[4..8]=s[10..14]$

$r[2..6]=s[5..9]$

$r[3..7]=s[6..10]$



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



FASTA file format

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

;a simple sequence in FASTA format **Comment**

Start sign > 2HBS:A|PDBID|CHAIN|SEQUENCE **Sequence ID**

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFL
SFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTN
AVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLS
HCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL
TSKYR **Sequence**

- 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*: CCTACCGATAACCGA
- 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 <i>r</i>	Position	5-mer in <i>s</i>	Position in <i>s</i>	Corr. Match in <i>r</i>	Matched 5-mer in <i>r</i>
ACCGA	4, 10	ACCTA	1	4, 10	ACCGA
ATACC	8	CCTAT	2	5, 1	CCGAT, CCTAC
CCGAT	5	CTATA	3	6	CGATA
CCTAC	1	TATAG	4		
CGATA	6	ATAGC	5	8	ATACC
CTACC	2				
GATAC	7				
TACCG	3, 9				

Answer: BLAST

- Extend local matches to longer matches:

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

r: 4 ACCGATA 10
s: 1 ACCTATA 7

r: 8 ATACC 12
s: 5 ATAGC 9

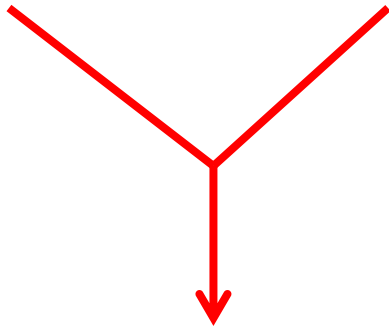
r: 10 ACCGA 14
s: 1 ACCTA 5

r: 1 CCTAC 5
s: 2 CCTAT 6

Answer: BLAST

- Combine matches and calculate scores:

r: 4 ACCGATA 10 8 ATACC 12
s: 1 ACCTATA 7 5 ATAGC 9



r: 4 ACCGATACC 12
s: 1 ACCTATAGC 9

Score = 7

10 ACCGA 14
1 ACCTA 5



10 ACCGA 14
1 ACCTA 5

Score = 4

1 CCTAC 5
2 CCTAT 6



1 CCTAC 5
2 CCTAT 6

Score = 4

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 <i>s</i> in the database	Number of matches in each pair	Number of pairs
2 “GC” in <i>s</i>	5’ -GCGC-3’	2	1
1 “GC” in <i>s</i>	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 <i>s</i> in the database	Number of matches in each pair	Number of pairs
4 “A” in <i>s</i>	5’ -AAAA-3’	3	1
3 “A” in <i>s</i>	5’ -AAAX-3’	2	3
	5’ -XAAA-3’		3
2 “A” in <i>s</i>	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:

number of matches with score ≥ 2 (exact match in this case)

total number of possible combinations

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

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

$$= 0.18164$$

Comparison between FASTA and BLAST

	FASTA	BLAST
Full name	FAST for All kinds of sequences	Basic Local Alignment Search Tool
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

		Second base					
		U	C	A	G		
First base	U	UUU } Phenylalanine F UUC UUA } Leucine L UUG	UCU } Serine S UCC UCA UCG	UAU } Tyrosine Y UAC UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC UGA } Stop codon UGG } Tryptophan W	Third base	U
	C	CUU } Leucine L CUC CUA CUG	CCU } Proline P CCC CCA CCG	CAU } Histidine H CAC CAA } Glutamine Q CAG	CGU } Arginine R CGC CGA CGG		C
	A	AUU } Isoleucine I AUC AUA AUG } Methionine start codon M	ACU } Threonine T ACC ACA ACG	AAU } Asparagine N AAC AAA } Lysine K AAG	AGU } Serine S AGC AGA } Arginine R AGG		A
	G	GUU } Valine V GUC GUA GUG	GCU } Alanine A GCC GCA GCG	GAU } Aspartic acid D GAC GAA } Glutamic acid E GAG	GGU } Glycine G GGC GGA GGG		G

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

The screenshot shows the NCBI BLAST website. The browser address bar displays `blast.ncbi.nlm.nih.gov/Blast.cgi`. The page header includes the NIH logo, "U.S. National Library of Medicine", "NCBI National Center for Biotechnology Information", and a "Sign in to NCBI" link. The main heading is "BLAST®" with navigation links for "Home", "Recent Results", "Saved Strategies", and "Help".

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](#)

Web BLAST

The "Web BLAST" section features four options, each in a colored box with a molecular diagram:

- Nucleotide BLAST** (teal box, red border): nucleotide ► nucleotide. This option is highlighted with a red rectangular border.
- blastx** (teal box): translated nucleotide ► protein.
- tblastn** (teal box): protein ► translated nucleotide.
- Protein BLAST** (blue box): protein ► protein.

BLAST Genomes

Search Betacoronavirus Database

We have created a new BLAST database focused on the SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences. For further detail please visit [NCBI GenBank](#).

Mon, 03 Feb 2020 10:00:00 EST [More BLAST news...](#)

Step 2: Input the sequence in FASTA format

>UNKNOWN_GENE

ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT

Nucleotide BLAST: Search number: X

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® » blastn suite Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

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

>UNKNOWN_GENE
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT

Query subrange From To

BLAST results will be displayed in a new format by default. You can always switch back to the Traditional Results page.

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

Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id—completions will be suggested ☐ exclude 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

Entrez Query Optional Enter an Entrez query to limit search [You Labs](#) [Create custom database](#)

Program Selection

Optimize for ☒ Highly similar sequences (megablast) ☐ More dissimilar sequences (discontinuous megablast)

Step 3: Set the parameters

The screenshot shows the NCBI Nucleotide BLAST search page. The browser address bar displays the URL: `blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome`.

Enter Query Sequence

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

>UNKNOWN_GENE
ATGGTGATCTGACTCCTGAGGAGAAGTCTGCCGTACT

Query subrange
From
To

BLAST results will be displayed in a new format by default
You can always switch back to the Traditional Results page.

Or, upload file No file chosen

Job Title
UNKNOWN_GENE
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
Human genomic plus transcript (Human G+T)

Exclude
Optional
☐ Sequences from type material

Limit to
Optional
Enter an Entrez query to limit search

Entrez Query
Optional
[YouTube](#) [Create custom database](#)

Program Selection

Optimize for
☒ Highly similar sequences (megablast)
☐ Moderately similar sequences (discontiguous megablast)
☐ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST Search database Human G+T using Megablast (Optimize for highly similar sequences)
☐ Show results in a new window

[+ Algorithm parameters](#)

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

Step 4: Show the results

Job Title UNKNOWN_GENE

RID 4G30UWUT014 *Search expires on 02-16 22:19 pm* [Download All](#) ▼

Program BLASTN [Citation](#) ▼

Database GPIPE/9606/current/ref_top_level GPIPE/9606/current/rna (2 databases) [See details](#) ▼

Query ID lcl|Query_47109

Description UNKNOWN_GENE

Molecule type dna

Query Length 39

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 ▼ Manage Columns ▼ Show 100 ▼ ?

☒ select all 2 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Transcripts							
☒	Homo sapiens hemoglobin subunit beta (HBB) mRNA	73.1	73.1	100%	5e-12	100.00%	NM_000518.4
Genomic sequences [show first]							
☒	Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly	73.1	73.1	100%	5e-12	100.00%	NC_000011.10

[Feedback](#)

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
```

The screenshot shows a web-based alignment viewer interface. At the top, there are tabs for 'Descriptions', 'Graphic Summary', 'Alignments' (which is selected), and 'Taxonomy'. Below the tabs, the 'Alignment view' is set to 'Pairwise'. A status bar indicates '2 sequences selected'. Below this, there are links for 'Download', 'GenBank', and 'Graphics'. The main content area displays the alignment details for 'Homo sapiens hemoglobin subunit beta (HBB), mRNA' with sequence ID 'NM_000518.4'. It shows a range of 51 to 89 and 1 match. A table provides summary statistics: Score of 73.1 bits(39), Expect of 5e-12, Identities of 39/39(100%), Gaps of 0/39(0%), and Strand of Plus/Plus. At the bottom, the sequence alignment is shown with vertical bars indicating matches between the query and subject sequences.

Score	Expect	Identities	Gaps	Strand
73.1 bits(39)	5e-12	39/39(100%)	0/39(0%)	Plus/Plus

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

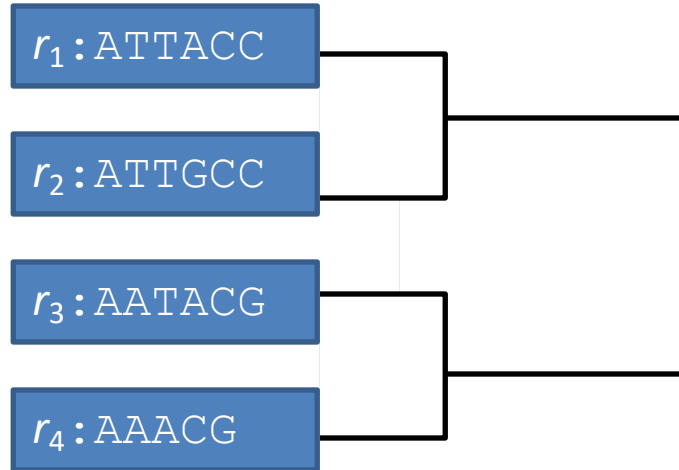
Multiple sequence alignment (MSA)

- 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)

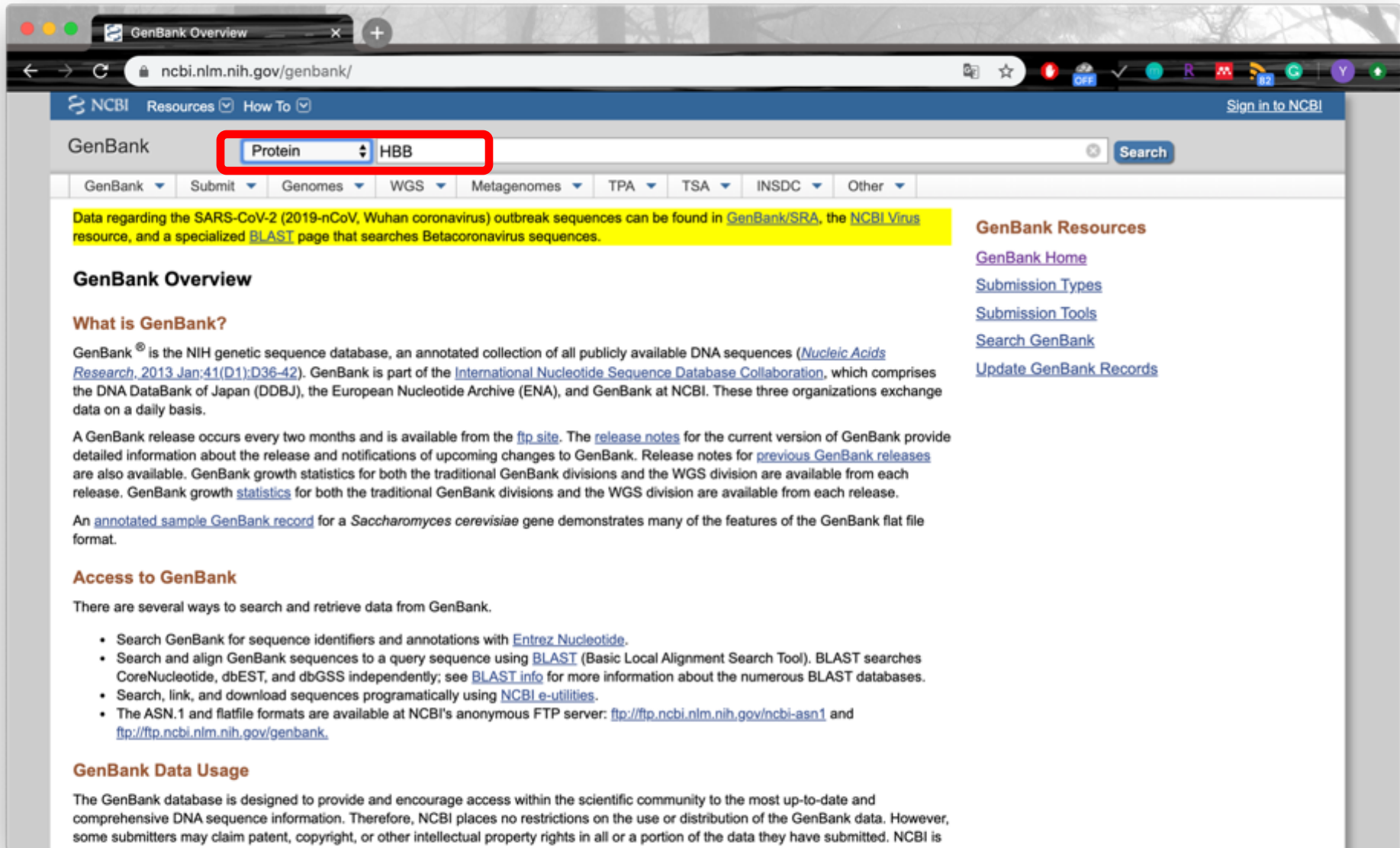
Possible alignment:

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”



The screenshot shows the GenBank Overview page on the NCBI website. The browser address bar displays ncbi.nlm.nih.gov/genbank/. The search bar is highlighted with a red rectangle, showing the search type set to "Protein" and the search term "HBB". Below the search bar, there are navigation tabs for GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, and Other. A yellow banner provides information about SARS-CoV-2 sequences. The main content area is titled "GenBank Overview" and includes a section "What is GenBank?" which describes the database and its history. To the right, there is a "GenBank Resources" sidebar with links to GenBank Home, Submission Types, Submission Tools, Search GenBank, and Update GenBank Records. At the bottom, there is a section "Access to GenBank" with a list of search methods and a "GenBank Data Usage" section.

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research](#), 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utils](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is

Get HBB protein sequences from GenBank

NCBI Protein search results for HBB. The page displays various filters and search options. A red box highlights the 'Results by taxon' section, which lists top organisms and their counts.

Results by taxon

- Top Organisms [Tree]
- Homo sapiens (1079)
- Peromyscus maniculatus (347)
- Mus musculus (171)
- Pan troglodytes (161)
- Hepatitis B virus (160)
- All other taxa (7503)
- More...

RefSeq Sequences

Items: 1 to 20 of 9421

1. **HBB, partial [synthetic construct]**
147 aa protein
Accession: AKI70611.1 GI: 823670800
[Nucleotide](#) [PubMed](#) [Taxonomy](#)

Take Homo Sapiens as an example ...

NCBI Resources How To Sign in to NCBI

Protein (HBB) AND "Homo sapiens"[porgn:txid9606] Search

Create alert Advanced Help

Species Animals (1,079) Customize ...

Source databases PDB (8) RefSeq (11) UniProtKB / Swiss-Prot (8) Customize ...

Sequence length Custom range...

Molecular weight Custom range...

Release date Custom range...

Revision date Custom range...

Clear all Show additional filters

Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

See the results of this search (1 item) in our new Identical Protein Groups database.

Items: 1 to 20 of 1079

<< First < Prev Page 1 of 54 Next > Last >>

☐ [HBB, partial \[Homo sapiens\]](#)
1. 147 aa protein
Accession: CAG46711.1 GI: 49456781
[Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ [HBB \[Homo sapiens\]](#)
2. 147 aa protein
Accession: CAG38767.1 GI: 49168544
[Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ [RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain; Contains: RecName: Full=LVV-hemorphin-7; Contains: RecName: Full=Spinorphin](#)
3. 147 aa protein
Accession: P68871.2 GI: 56749856
[PubMed](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ [hemoglobin subunit beta \[Homo sapiens\]](#)
4. 147 aa protein
Accession: NP_000509.1 GI: 4504349
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

Find related data
Database: Select
Find items

Search details
HBB[All Fields] AND "Homo sapiens"[porgn]
Search See more...

Recent activity
Turn Off Clear
Q (HBB) AND "Homo sapiens"[porgn] (1079) Protein
Q HBB (9421) Protein
UK biobank: an open access resource for identifying the causes of a wide range PubMed
R PheWAS: data analysis and plotting tools for phenome-wide association studies PubMed
Phenome-wide association studies on a

Take Homo Sapiens as an example ...

ncbi.nlm.nih.gov/protein/CAG46711.1

NCBI Resources How To Sign In to NCBI

Protein Protein Search Advanced Help

GenPept

HBB, partial [Homo sapiens]

GenBank: CAG46711.1

[Identical Proteins](#) **FASTA** [Graphics](#)

Go to: ▾

Send to: ▾ Change region shown ▾ Customize view ▾

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

Crystal structure of ferric R-state human methemoglobin bound to maleimide-PDB: 6HK2

Source: Homo sapiens

Method: X-ray Diffraction

Resolution: 1.55 Å

See all 234 structures...

Articles about the HBB gene

First Description of Hb San Diego (<i>HBB</i>: c.328G>A) in a Chinese Fi [Hemoglobin. 2019]

Computational Analysis of Protein Structure Changes as a Result of N [Biomed Res Int. 2019]

LOCUS CAG46711 147 aa linear PRI 26-JUL-2016

DEFINITION HBB, partial [Homo sapiens].

ACCESSION CAG46711

VERSION CAG46711.1

DBSOURCE embl accession [CR541913.1](#)

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 147)

AUTHORS Halleck,A., Ebert,L., Mkoundinya,M., Schick,M., Eisenstein,S., Neubert,P., Kstrang,K., Schatten,R., Shen,B., Henze,S., Mar,W., Korn,B., Zuo,D., Hu,Y. and LaBaer,J.

TITLE Cloning of human full open reading frames in Gateway(TM) system entry vector (pDONR201)

JOURNAL Unpublished

REFERENCE 2 (residues 1 to 147)

AUTHORS Halleck,A., Ebert,L., Mkoundinya,M., Schick,M., Eisenstein,S., Neubert,P., Kstrang,K., Schatten,R., Shen,B., Henze,S., Mar,W., Korn,B., Zuo,D., Hu,Y. and LaBaer,J.

TITLE Direct Submission

JOURNAL Submitted (28-JUN-2004) RZPD Deutsches Ressourcenzentrum fuer Genomforschung GmbH, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany

COMMENT RZPD: RZPD:824506320, OPEN: 3042

Take Homo Sapiens as an example ...

The screenshot shows the NCBI protein database entry for HBB, partial [Homo sapiens] (GenBank: CAG46711.1). The FASTA sequence is highlighted with a red box:

```
>CAG46711.1 HBB, partial [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAMGNPKVKAHGKKVLG
AFSDGLAHLNLRKGTATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

The right sidebar contains several sections:

- Analyze this sequence**
 - Run BLAST
 - Identify Conserved Domains
 - Highlight Sequence Features
 - Find in this Sequence
- Protein 3D Structure**
 - Crystal structure of ferric R-state human methemoglobin bound to maleimide-PDB: 6HK2
 - Source: Homo sapiens
 - Method: X-ray Diffraction
 - Resolution: 1.55 Å
 - See all 234 structures...
- Articles about the HBB gene**
 - First Description of Hb San Diego (<i>HBB</i>: c.328G>A) in a Chinese F₁ [Hemoglobin. 2019]
 - Computational Analysis of Protein Structure Changes as a Result of N [Biomed Res Int. 2019]
 - A Novel Human β -Globin Gene Variant [Hb London-Ontario, <i>HBB</i>: [Hemoglobin. 2019]

HBB protein sequences from different species

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

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLNLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

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

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLNLDNLKGTFATLSELHCDKLHVDPENFR

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

MVHLTDAEKAASGLWGKVNSDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAKVKAHGKKVIT
AFNEGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLGKDFTPAAQAAFQKVMAGVAT
ALAHKYH

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

MVHLTDAEKALVTGLWGKVKPEEIGGEALGRLLAVYPWTQRFFDSFGDLSSASAIMGNAKVKGHGKKVID
SFGEGLKHLNLDNLKGTFASLSELHCDKLHVDPENFKLLGNMIVIVMAHHLGKDFTPAAQAAYQKVVAGVAT
ALAHKYH

Find the tool for MSA: Clustal Omega

The screenshot shows a Google search interface with the query 'multiple sequence alignment'. The featured snippet from Wikipedia defines MSA as a sequence alignment of three or more biological sequences. Below it, a search result from EMBL-EBI for 'Clustal Omega' is highlighted with a red rectangle. The result title is 'Clustal Omega < Multiple Sequence Alignment < EMBL-EBI' and the description states it is a new MSA program using seeded guide trees and HMM profile-profile techniques. At the bottom, there is a section 'People also search for' with suggestions like 't coffee multiple sequence alignment ppt', 'boxshade multiple sequence alignment ncbi', and 'esript multiple sequence alignment similarity score'.

multiple sequence alignment

Google

multiple sequence alignment

Multiple alignment

- Multiple alignments are useful for comparing homologous sequences at once.
- Multiple alignment of part of Exon from different genes.
- Multiple alignments can be **global** or **local**.
- The purpose of multiple sequence alignment is to make multiple sequence alignment.
- In the sequence, there are several of high sequence similarity.
- Multiple alignment of part of Exon from different genes.
- You can "see" the region of similarity from each sequence.
- Multiple alignment of part of Exon from different genes.

A **multiple sequence alignment (MSA)** is a **sequence alignment** of three or more biological **sequences**, generally protein, DNA, or RNA. In many cases, the input set of query **sequences** are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor.

en.wikipedia.org › wiki › Multiple_sequence_alignment

[Multiple sequence alignment - Wikipedia](#)

About Featured Snippets Feedback

www.ebi.ac.uk › Tools › msa › clustalo

Clustal Omega < Multiple Sequence Alignment < EMBL-EBI

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments ...

People also search for

- t coffee multiple sequence alignment ppt
- boxshade multiple sequence alignment ncbi
- esript multiple sequence alignment similarity score

Input the sequences

Clustal Omega < Multiple Seq... X

ebi.ac.uk/Tools/msa/clustalo/

Clustal Omega

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>HumanlCAG46711.1 HBB, partial [Homo sapiens]
MVHLTPEEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDDLKGTFTATLSLHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

>ChimpanzeelACY75418.1 beta globin, partial [Pan troglodytes verus]
MVHLTPEEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDDLKGTFTATLSLHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Or, upload a file: No file chosen

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

Use the default parameters

The screenshot shows the Clustal Omega web interface. The browser address bar displays `ebi.ac.uk/Tools/msa/clustalo/`. The page has a teal header with navigation links: [Input form](#), [Web services](#), [Help & Documentation](#), [Bioinformatics Tools FAQ](#), [Feedback](#), and [Share](#).

The main content area is divided into two steps:

STEP 1 - Input your sequence

Sequence input: `>ChimpanzeeAC75418.1 beta globin, partial [Pan troglodytes verus]`
`MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLG`
`AESDCLAHLDNLIKKEATLSLHCKLVLDENED`

Buttons: [Choose File](#) (No file chosen), [Use a example sequence](#), [Clear sequence](#), [See more example inputs](#)

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUMBER of COMBINED ITERATIONS
no	yes	yes	default(0)
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER	
default	default	aligned	

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 this service, please consider citing the following publication: [The EMBL-EBI search and sequence analysis tools APIs in 2019](#)

Please read the provided [Help & Documentation](#) and [FAQs](#) before seeking help from our support staff. If you have any feedback or encountered any issues please let us

Show the alignment

Results < Clustal Omega < Multi

ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-l20200215-135553-0921-9086725-p1m

Clustal Omega

Input form

Web services

Help & Documentation

Bioinformatics Tools FAQ

Feedback

Share

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-l20200215-135553-0921-9086725-p1m

Alignments

Result Summary

Guide Tree

Phylogenetic Tree

Results Viewers

Submission Details

Download Alignment File

Show Colors

CLUSTAL O(1.2.4) multiple sequence alignment

Human|CAG46711.1

Chimpanzee|ACY75418.1

Mouse|BAG16710.1

Mouse|AJQ22215.1

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVHGNNPK
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVHGNNPK
MVHLTDAEKAASVGLWGVNSDEVGGEALGRLLVVYPWTQRFFESFGDLSSAIAINGNAK
MVHLTDAEKAALVTGLWGVNKPVEIGGEALGRLLAVYPTQRFDSFGDLSSAIAINGNAK
***** ** :*:***** :*:*****:*****: :**** *

60

Human|CAG46711.1

Chimpanzee|ACY75418.1

Mouse|BAG16710.1

Mouse|AJQ22215.1

VKANGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
VKANGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR-----
VKANGKKVITAFNEGLNHLDLSLGTFASSLSEHCDKLHVDPENFRLLGNMIVIVLGHHLG
VKGHGKKVIDSFGEGLKHLDNLKGTFASLSEHCDKLHVDPENFKLLGNMIVIMAAHHLG
.**: :*,:** ***.****:*****:

120

105

120

120

Human|CAG46711.1

Chimpanzee|ACY75418.1

Mouse|BAG16710.1

Mouse|AJQ22215.1

KEFTPPVQAAYQKVVAGVANALAHKYH

KDFTPAQAAPQKVMAGVATALAHKYH
KDFTPAQAAYQKVVAGVATALAHKYH

147

105

147

147

PLEASE NOTE: Showing colors on large alignments is slow.

Waiting for www.ebi.ac.uk...

Show the tree

The screenshot shows the EBI Clustal Omega web interface. The browser address bar displays the URL: ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-l20200215-135553-0921-9086725-p1m&a.... The page has a teal header with navigation links: Input form, Web services, Help & Documentation, Bioinformatics Tools FAQ, Feedback, and Share. Below the header, the breadcrumb trail reads: Tools > Multiple Sequence Alignment > Clustal Omega. The main heading is "Results for job clustalo-l20200215-135553-0921-9086725-p1m". A row of tabs includes: Alignments, Result Summary, Guide Tree, Phylogenetic Tree (selected), Results Viewers, and Submission Details. Below the tabs is a button labeled "Download Phylogenetic Tree Data". The section title "Phylogenetic Tree" is followed by the text: "This is a Neighbour-joining tree without distance corrections." Below this, there are two radio buttons for "Branch length": "Cladogram" (selected) and "Real". To the left of the text is a simple cladogram showing a root branching into two main groups. The first group branches into Human and Chimpanzee. The second group branches into Mouse and another Mouse entry. To the right of the cladogram, the following sequence identifiers and distances are listed:
Human|CAG46711.1 0.0034
Chimpanzee|ACY75418.1 -0.0034
Mouse|BAG16710.1 0.04762
Mouse|AJQ22215.1 0.08163

Below the tree visualization is the section "Tree Data" containing a Newick format string:

```
{ Human|CAG46711.1:0.00340,  
Chimpanzee|ACY75418.1:-0.00340,  
{ Mouse|BAG16710.1:0.04762,  
Mouse|AJQ22215.1:0.08163}  
:0.15306};
```

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?