



Foundation of Bio-Informatics

Lecture 3

Lecture by: Ahmad R. Naghsh-Nilchi, PhD

Department of Artificial Intelligence

Faculty of Computer Engineering

University of Isfahan



Going from Protein to DNA

- Correspondence between Protein and DNA sequences is not one-to-one.
- Many different DNA sequences can be linked to the same protein or gene.
- Some situations require going back to the DNA sequence such as trying to amplify a gene to transfer to another organism, a process called PCR.
- Given a protein sequence, *how can a correct DNA sequence be retrieved from a protein sequence?*



Retrieval Process Example: Hemoglobin

- Understanding Proteins' genetic foundation is critical in molecular biology, genetics, and biochemistry.
- For example, **hemoglobin**, a protein in red blood cells, is responsible for transporting oxygen from the lungs to the tissues and facilitating the return of carbon dioxide from the tissues to the lungs.
- Scientists use genetic information that explains biochemical pathways and their physiological roles to enhance knowledge of hemoglobin's function and its implications in clinical research, such as cell disease management, i.e. thalassemia management.



Step 1: Searching Protein Sequence in Databases



- NCBI or UniProt provides amino acid sequences.
- Hemoglobin has various forms, including human hemoglobin (HbA), characterized by its composition of two alpha and two beta subunits.
- Using the UniProt database, input "hemoglobin" in its search bar. A list of hemoglobin proteins across various organisms is given. Looking for a specific identifier, UniProt **accession number**.
- For instance, the accession number for human hemoglobin subunit beta is "P68871."



Searching UniProt for "hemoglobin"



UniProt [BLAST](#) [Align](#) [Peptide search](#) [ID mapping](#) [SPARQL](#) [UniProtKB](#) - hemoglobin [Advanced](#) | [List](#) [Search](#)

Status
 Reviewed (Swiss-Prot) (1,482)
 Unreviewed (TrEMBL) (68,794)

Popular organisms
Human (215)
Rat (114)
Mouse (107)
Zebrafish (71)
Bovine (60)

Taxonomy
[Filter by taxonomy](#)

Group by
[Taxonomy](#)
[Keywords](#)
[Gene Ontology](#)
[Enzyme Class](#)

Proteins with

UniProtKB 70,276 results

or search "hemoglobin" as a Protein Name, Gene Ontology, Protein family, Catalytic Activity, Disease, or Gene Name

[Tools](#) [Download \(70k\)](#) [Add](#) View: [Cards](#) [Table](#) [Customize columns](#) [Share](#)

| Entry | Entry Name | Protein Names | Gene Names | Organism | Length |
|---|----------------------------|---------------------------------|--|---|--------|
| <input type="checkbox"/> P68871 | HBB_HUMAN | Hemoglobin subunit beta[...] | HBB | Homo sapiens (Human) | 147 AA |
| <input type="checkbox"/> P02100 | HBE_HUMAN | Hemoglobin subunit epsilon[...] | HBE1 , HBE | Homo sapiens (Human) | 147 AA |
| <input type="checkbox"/> P02042 | HBD_HUMAN | Hemoglobin subunit delta[...] | HBD | Homo sapiens (Human) | 147 AA |
| <input type="checkbox"/> P02008 | HBAZ_HUMAN | Hemoglobin subunit zeta[...] | HBZ , HBZ2 | Homo sapiens (Human) | 142 AA |
| <input type="checkbox"/> P09105 | HBAT_HUMAN | Hemoglobin subunit theta-1[...] | HBQ1 | Homo sapiens (Human) | 142 AA |
| <input type="checkbox"/> P69891 | HBG1_HUMAN | Hemoglobin subunit gamma-1[...] | HBG1 , PRO2979 | Homo sapiens (Human) | 147 AA |
| <input type="checkbox"/> P69892 | HBG2_HUMAN | Hemoglobin subunit gamma-2[...] | HBG2 | Homo sapiens (Human) | 147 AA |
| <input type="checkbox"/> P11517 | HBB2_RAT | Hemoglobin subunit beta-2[...] | | Rattus norvegicus (Rat) | 147 AA |
| <input type="checkbox"/> P69905 | HBA_HUMAN | Hemoglobin subunit alpha[...] | HBA1 , HBA2 | Homo sapiens (Human) | 142 AA |
| <input type="checkbox"/> P02091 | HBB1_RAT | Hemoglobin subunit beta-1[...] | Hbb | Rattus norvegicus (Rat) | 147 AA |
| <input type="checkbox"/> P01966 | HBA_BOVIN | Hemoglobin subunit alpha[...] | HBA | Bos taurus (Bovine) | 142 AA |
| <input type="checkbox"/> P83479 | HBAC_CONGO | Hemoglobin cathodic subunit | | Conger conger (Conger eel) (Muraena conger) | 143 AA |



Step 2: Retrieve the Corresponding Gene ID and Details



- Gene identifier (ID) provides deeper insights into a protein's genomic context.
- For hemoglobin, the gene responsible for encoding the beta subunit is termed "HBB," located on chromosome 11 of the human genome.
- With the UniProt "P68871" hemoglobin beta subunit in hand, we can use the NCBI site (or go directly to search NCBI's Gene database), where the "HBB" gene ID (e.g., 3043 in NCBI) is detailed.



Searching NCBI for Gene ID



NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI

P68871 gene id

Search

Results found in 7 databases

| Literature | Genes |
|--------------------|----------------|
| Bookshelf 0 | Gene 2 |
| MeSH 0 | GEO DataSets 0 |
| NLM Catalog 0 | GEO Profiles 0 |
| PubMed 0 | PopSet 0 |
| PubMed Central 127 | |

| Genomes | Clinical |
|---------------------------------|----------------------|
| Assembly / Genome NCBI Datasets | ClinicalTrials.gov 0 |
| BioCollections 0 | ClinVar 22 |
| BioProject 1,033 | dbGaP 0 |
| BioSample 12,089 | dbSNP 0 |
| Nucleotide 0 | dbVar 0 |
| SRA 0 | GTR 2 |
| Taxonomy 0 | MedGen 0 |
| | OMIM 0 |

NCBI Datasets All assembled genome data is now available through NCBI Datasets

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Gene

Gene P68871 Gene ID

Create RSS Save search Advanced Help

Tabular Sort by Relevance Send to:

Hide sidebar >>

Filters: Manage Filters

Find related data

Database: Select

Find items

Search details

(P68871[All Fields] AND ID[All Fields]) AND alive[prop]

Search See more...

Recent activity

Turn Off Clear

P68871 Gene ID AND alive[prop] (2)

Gene

See ID1 (ID) inhibitor of DNA binding 1 in the Gene database
id in [Homo sapiens](#) [Mus musculus](#) [Oryza sativa Japonica Group](#) [All 4](#)
[Gene records](#)

Search results

Items: 2

Showing Current items.

| Name/Gene ID | Description | Location | Aliases | MIM |
|---|---|--|------------------------------|--------|
| <input type="checkbox"/> HBB ID: 3043 | hemoglobin subunit beta [<i>Homo sapiens</i> (human)] | Chromosome 11, NC_000011.10 (5225484..5227071, complement) | CD113t-C, ECTY6, beta-globin | 141900 |
| <input type="checkbox"/> HBG2 ID: 3048 | hemoglobin subunit gamma 2 [<i>Homo sapiens</i> (human)] | Chromosome 11, NC_000011.10 (5253188..5254781, complement) | HBG-T1, TNCY | 142250 |



Gene Identity (ID):

- This ID serves as a critical reference point for further inquiries, granting access to a wealth of additional genetic information, such as gene structure, function, mutation data, and associated pathologies.



Step 3: Use Bioinformatics Tools to Obtain the DNA Sequence



- The NCBI provides several resources, including the GenBank database, which archives the sequences of genes obtained from various species.
- Gene page associated with HBB on the NCBI can locate a prominent feature that details the genomic sequence, including the full-length cDNA (complementary DNA) sequence.
- The results indicate that the HBB gene has a specific coding region composed of exons and introns. The mRNA sequence transcribed from this gene corresponds to the hemoglobin beta-subunit amino acid chain.



Searching GenBank

 **National Library of Medicine**
National Center for Biotechnology Information

Log in

GenBank

Nucleotide ▼

3043

Search

GenBank ▼ Submit ▼ Genomes ▼ WGS ▼ Metagenomes ▼ TPA ▼ TSA ▼ INSOC ▼ Documentation ▼ Other ▼

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 Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)



This tool allows downloading the entire genomic sequence directly in FASTA or other appropriate formats.



Nucleotide

Nucleotide

3043

Advanced

Search

GenBank

N. crassa NUO-32 mRNA for NADH dehydrogenase 32 kD subunit

GenBank: X56237.1

[FASTA](#) [Graphics](#)

Go to: ☺

LOCUS X56237 1077 bp mRNA linear PLN 12-JUN-2006
DEFINITION N. crassa NUO-32 mRNA for NADH dehydrogenase 32 kD subunit.
ACCESSION X56237
VERSION X56237.1
KEYWORDS NADH dehydrogenase (ubiquinone); NADH dehydrogenase subunit; NUO-32 gene.
SOURCE Neurospora crassa
ORGANISM Neurospora crassa
Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;

Nucleotide

Nucleotide

Advanced

Search

Help

FASTA

Send to: -

Change region shown

N. crassa NUO-32 mRNA for NADH dehydrogenase 32 kD subunit

GenBank: X56237.1

[GenBank](#) [Graphics](#)

>X56237.1 N. crassa NUO-32 mRNA for NADH dehydrogenase 32 kD subunit
CCACAATCCACCCACCCACCGAGTCACTCAACACCTCCACGACAGACTCGGAAACCATG
CGCGGGCACTCCGATTGCTCGCGACCGCGACGGCTACGTCGCCCGTCCGCCGCTTCTCAAACCCG
GCTCCCGACCGGCTCAACGGCTCGGCACCCACCGTCCGCCGCTCCGCCGCTGCTACCTCTACAA
CCACACCTCGACAAGCTCAAGCAGATCCCCGAGCACTCGCTGTACCGCAGTCCGCCGAGGCCCTGACC
AAGCACCGCTCGCCATCGTCGAGCAGTACGTGCCGACGGCTACGACGCTGGCAGGAGCGGCCGCA
AGCTGCTCGAGAAGCACAAGAGCGACCTCACGGCCCGCAGTTCGACGGCCAGCATGCCGCCAGTCTGA
AGGCCCGACGGCCGCTTACTTCATCCGCCAGATGGTCCCCCGCAGGACTGGCGCGACGTGAGTGG
GATGGTGGCTCCTGGATCCGCACTTTTCTGGGTTAGACCGGCGAGGATGTCGTCGGCGCCGTCAAGC
TGAAGACAGCGACAAGCTGCTCGAGCTGGACAAGATCAGGGAATCAGACCCGGTGCCTATCGCCAGGG
TCTCAGGGATCTCGGTATCAAGATGGGCGGTGTTGTCGAGGACAAGAGCCCGTCCGAGTGGGAGTCGGAG
CCGCCGTTGAGCGTGAACAGATTGCTGAGATGGAGGCGAGGATTGGTTCTGGCTTGATTGAGGAGGTTG
TTCAGGTTGCCGAGGGCGAGCTTAAGCTGGTTGACATTATGACCCAGGCGAGGCTTGGGAAGCTCTTGA
GGAGGAGGCGCCAGAGGGCAATGGACTTACTTTGAGCGTAAGGAGTAAACAAACATCGACAGACATAGA
GAATGGGAACACAGGACCAACCGGCAATGACGGCCACTATTTATGAGCGTTACGAAGAAGAAGGGAG
GACAGGCAATTTCTCAGCCAGTTGACCTATACCATACCGCGGACATGTATATAATCATCGACTTTG
CGAGGCGGGAATGACTATGCCCTCCTT

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Related information

Protein

PubMed

Taxonomy

Functional Class

PubMed (Weighted)

Recent activity



Comparing Protein Sequences

- The analysis of protein sequences is a fundamental aspect of molecular biology and bioinformatics, as it provides insights into the evolutionary relationships among proteins, the functional characteristics of proteins, and the mechanisms of various biological processes.
- Basic Local Alignment Search Tool (BLAST) is a vital tool for comparing protein sequences.



Understanding BLAST: An Overview

- BLAST, born in 1990, is an algorithm that identifies regions of similarity between biological sequences.
- BLAST can compare nucleotide or amino acid sequences,
- BLAST mission is much bigger than this, talk over it later.
- BLAST operates by analyzing the query sequence against a database of known sequences.
- BLAST algorithm segments sequences that yield significant matches based on a scoring system that takes into account several parameters.
- BLAST search returns a ranked list of database entries that have significant similarity to the query sequence.



The Steps to Conduct a BLAST Search for Protein Sequences



- 1. **Preparing the Query Sequence:**
- The initial step in using BLAST is to prepare a query sequence.
- This query can be derived from any source, including experimental data, theoretical models, or previously published sequences.
- The sequence must be in a standard format FASTA, the universal format recognized by BLAST tools.



2. Choosing the Appropriate BLAST Program:

- Depending on the nature of the sequences to be compared, researchers must select the appropriate variant of the BLAST algorithm.
- The principal variants include BLASTP for comparing protein sequences against proteins,
- BLASTX for translating nucleotide sequences and comparing them against protein sequences,
- PSI-BLAST and DELTA-BLAST use iterative methods to refine search results by focusing on previously identified homologs.



3. Selecting a Database

- 3. The database against which the query sequence will be compared is a critical consideration.
- BLAST offers several curated databases, including **GenBank**, **Swiss-Prot**, and specific organism databases,
- Users can upload custom databases.
- The choice of the database can significantly affect the results,
- Select databases relevant to your research context.



4. Configuring the Search Parameters

- Users can customize search parameters, such as the scoring matrix (e.g., BLOSUM62 or PAM), gap penalties, and limits on the number of alignments returned.
- Configuring these parameters can enhance the sensitivity and specificity of the search, thus producing more relevant results.
- These parameters are discussed later.



5. Running the BLAST Search

- Following configuration, BLAST search is executed.
- The algorithm processes the query against the selected database, identifying **High Scoring Pairs (HSPs)** and calculating scores and statistical significance using the “**Expect value**” (E-value).
- E-value represents the number of matches one expects to see by chance in a database of given size.
- Thus, lower E-values indicate more statistically significant results.



6. Interpreting the Results:

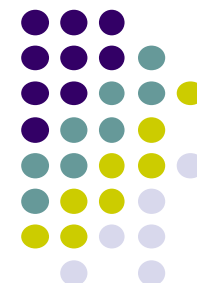


- Results of a BLAST search typically include an array of top hits, each accompanied by metadata such as
 - sequence identity, alignment length,
 - E-value, query cover,
 - and the associated taxonomic information.
- One must interpret these results in the context of his/her original biological questions.
- For instance, a high degree of sequence similarity may suggest evolutionary relatedness or functional similarity. (Caution is required! Other factors can complicate the interpretation)



Run BLAST by an Example

- Consider analyzing the similarities between the **human hemoglobin** protein and its counterparts in other species, such as chimpanzees and mice.
- First, the amino acid sequence of human hemoglobin is retrieved from NCBI databases, the previous example:
 - “P68871” hemoglobin, "HBB" gene ID 3043 in NCBI
- This sequence serves as the query for the BLAST search.



RUN BLAST on NCBI



BLAST® » blastp suite

Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query.

Enter Query Sequence

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

CAA40828.1

Query subrange [?](#)

From

To

Or, upload file

[Choose File](#)

No file chosen [?](#)

Job Title

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

☐ Align two or more sequences [?](#)

Choose Search Set

Databases

☒ Standard databases (nr etc.): ☐ Experimental databases

Compare

☐ Select to compare standard and experimental database [?](#)

Standard

Database

Non-redundant protein sequences (nr) [?](#)

Organism
Optional

Enter organism name or id—completions will be suggested

☐

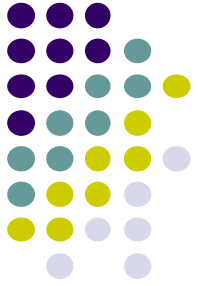
exclude

[Add organism](#)

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

Exclude
Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences



Utilize Appropriate Program:

- Utilizing the BLASTP program, which is specifically designed for comparing protein sequences, you should input the human hemoglobin sequence into the online interface.
- The program then compares the query sequence against its extensive protein database, assessing the potential alignments based on scoring metrics that reflect the degree of similarity.
- Results generated by BLAST include a list of sequences from other organisms that share homology with human hemoglobin.



Near the Bottom of the Page: Select BLASTP



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

Exclude
Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

- ☐ Quick BLASTP (Accelerated protein-protein BLAST)
☒ **blastp** (protein-protein BLAST)
☐ PSI-BLAST (Position-Specific Iterated BLAST)
☐ PHI-BLAST (Pattern Hit Initiated BLAST)
☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

BLAST

Search **database nr** using **Blastp** (protein-protein BLAST)

☒ Show results in a new window

+ Algorithm parameters

back

Format Request Status

[Formatting options]

Job Title: Human Hemoglobin Protein Comparison

| | |
|-----------------------|--------------------------|
| Request ID | MN9ARR30013 |
| Status | Searching |
| Submitted at | Sat Nov 30 00:40:39 2024 |
| Current time | Sat Nov 30 00:45:04 2024 |
| Time since submission | |

Descriptions

Graphic Summary

Alignments

Taxonomy

This page will be automatically updated in 30 seconds

Sequences producing significant alignments

Download ▾

Select columns ▾

Show 100 ▾



☒ select all 100 sequences selected

GenPept

Graphics

Distance tree of results

Multiple alignment

MSA Viewer

| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|-------------------------------------|--|---|-----------|-------------|-------------|---------|------------|----------|--------------------------------|
| <input checked="" type="checkbox"/> | NADH dehydrogenase (ubiquinone) 78 kDa subunit [Neurospora crassa] | Neurospora crassa | 1552 | 1552 | 100% | 0.0 | 100.00% | 744 | CAA40828.1 |
| <input checked="" type="checkbox"/> | NADH:ubiquinone oxidoreductase 78 [Neurospora crassa OR74A] | Neurospora crassa | 1476 | 1476 | 100% | 0.0 | 96.10% | 744 | XP_957188.3 |
| <input checked="" type="checkbox"/> | NADH dehydrogenase subunit [Neurospora crassa] | Neurospora crassa | 1475 | 1475 | 100% | 0.0 | 95.70% | 744 | AAA98999.1 |
| <input checked="" type="checkbox"/> | NADH-ubiquinone oxidoreductase 78 kDa subunit, mitochondrial [Neurospora hispaniola] | Neurospora hispaniola | 1473 | 1473 | 100% | 0.0 | 95.83% | 744 | XP_062691934.1 |
| <input checked="" type="checkbox"/> | NADH dehydrogenase 78K chain precursor [Neurospora tetrasperma FGSC 2508] | Neurospora tetrasperma | 1467 | 1467 | 100% | 0.0 | 95.43% | 744 | XP_009852144.1 |
| <input checked="" type="checkbox"/> | ndufs1 NADH-ubiquinone oxidoreductase subunit [Neurospora sp. IMI 360204] | Neurospora sp. IMI 360204 | 1463 | 1463 | 100% | 0.0 | 94.76% | 744 | KAJ4413690.1 |
| <input checked="" type="checkbox"/> | hypothetical protein QBC45DRAFT_418720 [Copromyces sp. CBS 386.78] | Copromyces sp. CBS 386.78 | 1453 | 1453 | 100% | 0.0 | 94.09% | 744 | KAK1776705.1 |
| <input checked="" type="checkbox"/> | hypothetical protein QBC32DRAFT_327037 [Pseudoneurospora amorphoporcata] | Pseudoneurospora amorphoporcata | 1451 | 1451 | 100% | 0.0 | 94.09% | 744 | KAK3949350.1 |
| <input checked="" type="checkbox"/> | hypothetical protein B0T09DRAFT_376610 [Sordaria sp. MPI-SDFR-AT-0083] | Sordaria sp. MPI-SDFR-AT-0083 | 1451 | 1451 | 100% | 0.0 | 93.82% | 744 | KAH7627016.1 |
| <input checked="" type="checkbox"/> | uncharacterized protein SMAC4_02450 [Sordaria macrospora] | Sordaria macrospora | 1450 | 1450 | 100% | 0.0 | 93.68% | 744 | XP_024511116.1 |
| <input checked="" type="checkbox"/> | hypothetical protein SMACR_02450 [Sordaria macrospora] | Sordaria macrospora | 1447 | 1447 | 100% | 0.0 | 93.55% | 744 | KAA8632318.1 |
| <input checked="" type="checkbox"/> | hypothetical protein B0T20DRAFT_253507 [Sordaria brevicollis] | Sordaria brevicollis | 1443 | 1443 | 100% | 0.0 | 93.28% | 744 | KAK3397274.1 |
| <input checked="" type="checkbox"/> | hypothetical protein C8A03DRAFT_15845 [Achaetomium macrosporum] | Achaetomium macrosporum | 1351 | 1351 | 100% | 0.0 | 86.96% | 743 | KAK4237643.1 |



Proteins Multiple Alignments

- The second most common bioinformatics task is multiple alignments.
- Multiple Alignments consist in lining up many similar proteins side-by-side for comparison.
- Details of multiple alignments are covered later.
- Now, just to show you that performing a multiple alignment is easy!



Multiple Alignments continued ...

- Multiple alignments are used to
 - Identify sequence positions where specific amino acids really matter for the structural integrity or the function of a given protein
 - Define specific sequence signatures for protein families
 - Classify sequences and build evolutionary trees.



Protein Information Resource (PIR)

- The Protein Information Resource (PIR) server is located at Georgetown University, helping to do multiple alignments.
- PIR originated from the **Atlas of Protein Sequences**, the first protein sequence collection (Prof. Dayhoff)
- PIR offers a multiple-alignment server, running the standard ClustalW program, that is easy to use.
- PIR available at: PIR.Georgetown.edu
- Now: <https://proteininformationresource.org/>



Step 1: Collection of Protein Sequences



- First step in a *protein multiple alignment* is the collection of sequences that are to be aligned.
- This is possible via databases: UniProt, NCBI, or specialized repositories like PIR, ExPAsY, Pfam.
- **Example:** consider aligning **three homologous protein sequences** from different species:
 - human hemoglobin (HBB_HUMAN),
 - mouse hemoglobin (HBB_MOUSE), and
 - cow hemoglobin (HBB_BOVIN).
- Their protein sequences can be retrieved in FASTA format like previous example.



One Word of Each Sequence:

- 1. Human Hemoglobin:
 - >HBB_HUMAN
 - MVHLTPEEKSAVTALWGKVNIA
 - ...
- 2. Mouse Hemoglobin:
 - >HBB_MOUSE
 - MVHLTPEEKSAVTALWGRVNIA
 - ...
- 3. Cow Hemoglobin:
 - >HBB_BOVIN
 - MVHLTPEEKSAVSALWGRVNIA
 - ...



Step 2: Preprocessing Sequences

- Once the sequences are collected, they must be preprocessed to ensure compatibility and accuracy during alignment. Common preprocessing steps include:
 - 1. **Cleaning sequences**: Removing any gaps or non-standard characters from the sequences.
 - 2. **Normalization**: Ensuring consistent formatting, such as upper or lower case, to avoid discrepancies.
 - 3. **Trimming**: Eliminating any non-informative regions or incomplete sequences.
- In our example, the sequences are already formatted uniformly with no apparent gaps or extra characters.



Step 3: Choosing an Alignment Method



- Several methods exist for protein multiple alignment. Common algorithms include:
 - **Progressive alignment:** This technique builds the alignment in stages, starting with the most similar sequences and progressively adding others.
 - **Iterative refinement:** This method iteratively improves the alignment based on predefined scoring matrices.
 - **Hidden Markov models (HMMs):** These statistical models can be useful for aligning sequences with complex evolutionary relationships.
- For our case, the progressive alignment method through software tool Clustal Omega, an effective tool in a variety of scenarios.



Step 4: Performing the Alignment

- Utilizing Clustal Omega, a brief outline of how to use the tool through the online platform:
 - 1. Navigate to the **Clustal Omega website**.
 - 2. **Paste the FASTA-formatted** sequences into the text box.
 - 3. Submit the sequences for alignment.
- The algorithm generates a multiple sequence alignment based on the protein sequences.
- The Clustal Omega tool employs a scoring system that takes into account substitution matrices such as BLOSUM62 or PAM, which help assess likelihood of amino acid substitutions.



Example Alignment Result



- The resulting alignment using UniProt:

UniProt Tools - UniProtKB - Advanced | List Search

Align

Find a protein sequence by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001) to align with the [Clustal Omega program](#).
You can also paste a list of IDs...

UniProt IDs

OR

Enter multiple protein or nucleotide sequences (50 max), separated by a FASTA header. You may also [load from a text file](#).

```
>HBB_HUMAN
MVHLTPEEKSAVTALWGKVNI
>HBB_MOUSE
MVHLTPEEKSAVTALWGRVNI
>HBB_BOVIN
MVHLTPEEKSAVSALWGRVNI
```

Your input contains 3 sequences.

Reset Align 3 sequences

Feedback

UniProt Tools - Tool results - Advanced | List Search

Align results

Overview Trees Percent Identity Matrix Text Output Input Parameters API Request

Tools - Download Add Resubmit

Highlight properties View: Continuous ☒ Wrapped

| | | | |
|-----------|---------------|-----------|---|
| HBB_BOVIN | MVHLTPEEKSAVS | SALWGRVNI | A |
| HBB_HUMAN | MVHLTPEEKSAVT | TALWGKVNI | A |
| HBB_MOUSE | MVHLTPEEKSAVT | TALWGRVNI | A |

22
22
22

Feedback



Step 5: Evaluating the Alignment

- Now, it is essential to evaluate the alignment accuracy and biological relevance.
- Several metrics can be used:
 - The consistency score, which measures how well the alignment agrees with known homologous relationships, and
 - The overall alignment score, which reflects the quality of the gaps and substitutions.

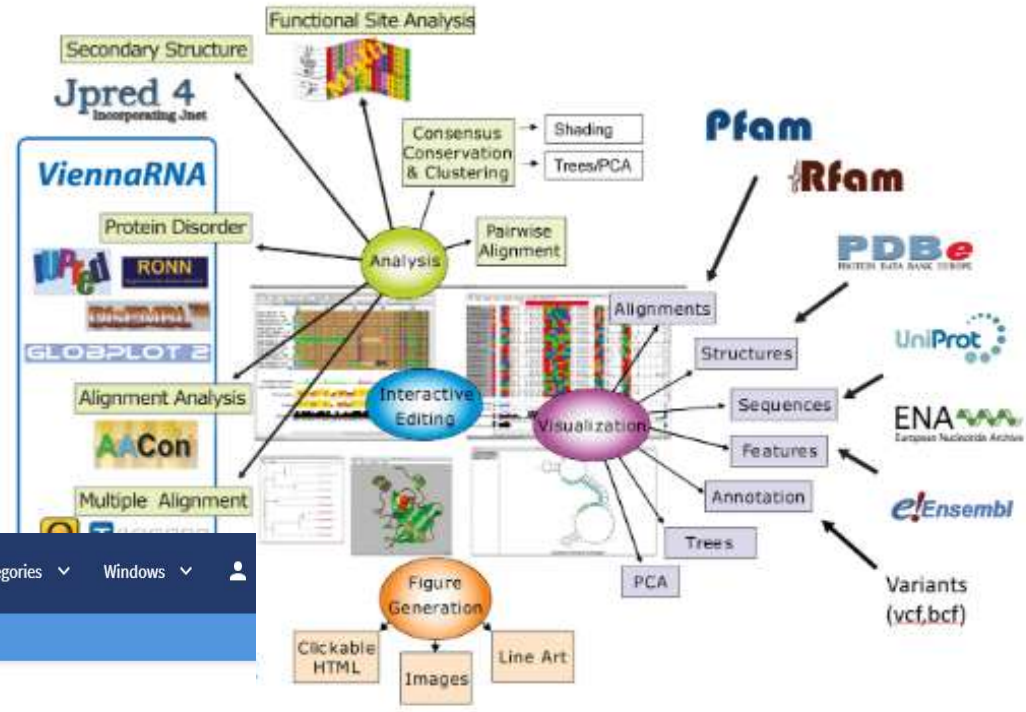


Step 5 Continued ...

- Tools such as **Jalview** or the **BioEdit** software can facilitate visual inspection of the alignment.
- This evaluation stage is crucial for identifying discrepancies or potential issues that need to be addressed.
- often necessitating manual adjustments or re-evaluations using different alignment algorithms.

[Home](#)[About](#)[Help](#)[Development](#)[Training](#)[Schools](#)[JalviewJS](#)[Download](#)

Jalview is a free cross-platform program for multiple sequence alignment editing, visualisation and analysis. Use it to align, view and edit sequence alignments, analyse them with phylogenetic trees and principal components analysis (PCA) plots and explore molecular structures and annotation.

[software.informer](#)[EN](#)[Categories](#)[Windows](#)[Windows](#) > [Education](#) > [Science](#) > [BioEdit](#)

BioEdit 7.7

FREE

Creates and edits biological sequence models

4 ★★★★★
797 votes

Latest version:
7.7.1 [See all](#)

Developer:
Tom Hall

Your vote:
★★★★★

Advertisement

[Download](#)**3 Easy Steps:**

1. Click "Download"
2. Start Download
3. Install the app

Get Fast!

[Review](#)[Download](#)[Comments \(34\)](#)[Questions & Answers \(24\)](#)[SHARE](#)



Step 6: Analyzing and Interpreting the Results



- Now one can proceed to analyze the results. Key objectives may include:
 - **Identifying conserved regions:** High conservation among sequences indicates crucial functional sites or domains.
 - **Characterizing evolutionary relationships:** By studying similarities and differences, one can infer phylogenetic relationships. This leads to insights regarding species evolution and divergence.
 - **Predicting functional implications:** Variations in the sequence can hint at functional diversity, allowing researchers to formulate hypotheses about protein activity or interactions.



Analyze our Example, Conclusion:



- In our example, one might identify a conserved active site essential for hemoglobin function, providing a basis for further exploration into evolution of oxygen transport mechanisms across species.
- Protein multiple alignment is an essential bioinformatics technique for understanding protein sequences and their evolutionary relationships.
- Systematic approach involving sequence collection, preprocessing, algorithm selection, execution, validation, and analysis ensures robustness and reliability in results.
- Through careful alignment and evaluation, one can illuminate essential biological insights that inform further scientific inquiry and discovery.

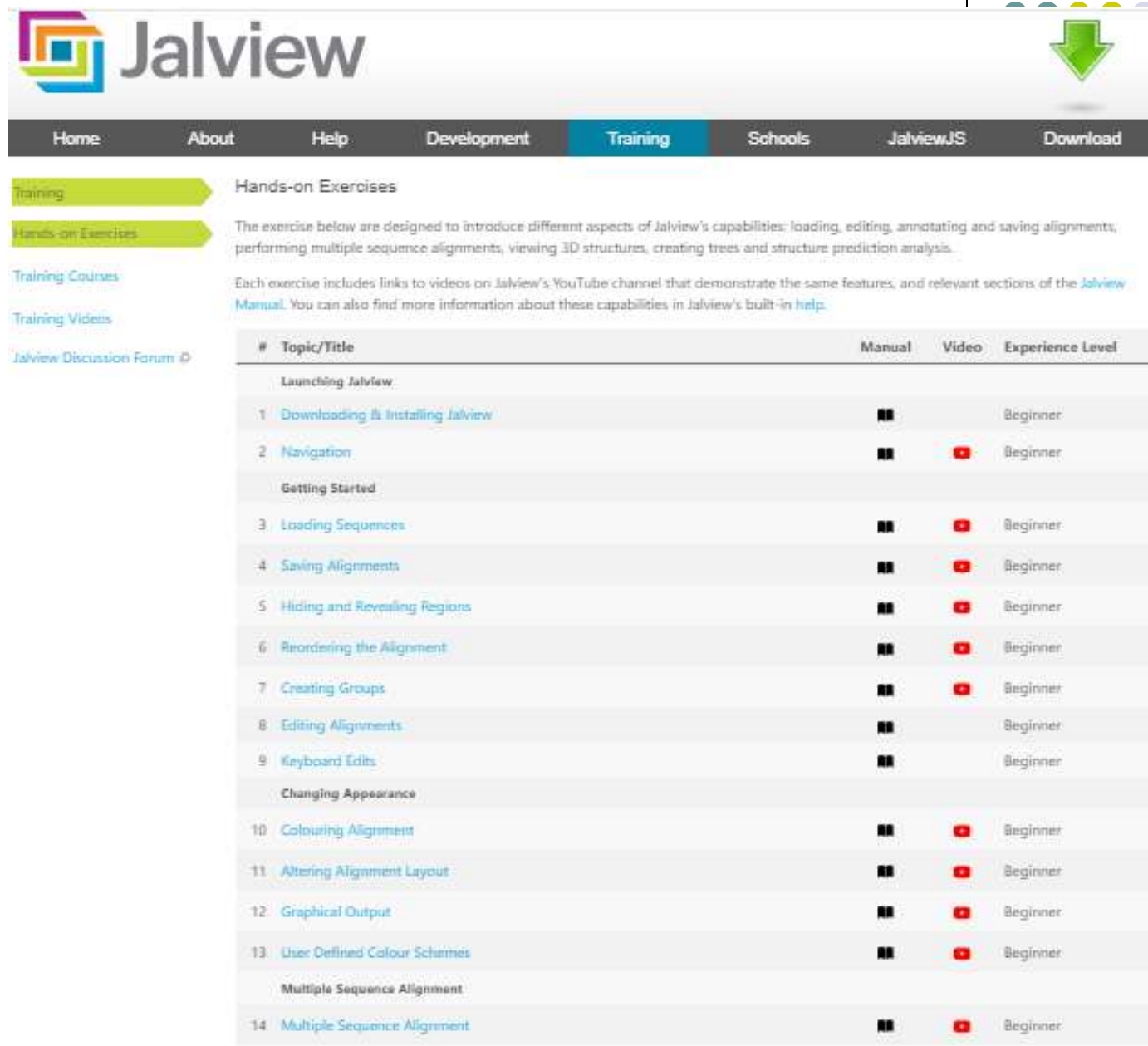


Home Exercise 1:

Visit “Training” section on Jalview and perform hands-on exercises 1-4 , 14.

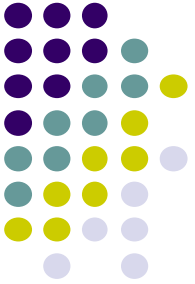
Write a report on your exercise. Print out your report (in Farsi) and hand it in next week.

Take the screen shut of your results and annex them into your report.



The screenshot shows the Jalview website's Training section. The navigation bar includes Home, About, Help, Development, Training (selected), Schools, JalviewJS, and Download. The Training section is titled "Hands-on Exercises" and includes a list of exercises with their respective manual and video links, and experience levels.

| # | Topic/Title | Manual | Video | Experience Level |
|------------------------------------|--|--------|-------|------------------|
| Launching Jalview | | | | |
| 1 | Downloading & Installing Jalview | ■ ■ | | Beginner |
| 2 | Navigation | ■ ■ | ■ | Beginner |
| Getting Started | | | | |
| 3 | Loading Sequences | ■ ■ | ■ | Beginner |
| 4 | Saving Alignments | ■ ■ | ■ | Beginner |
| 5 | Hiding and Revealing Regions | ■ ■ | ■ | Beginner |
| 6 | Reordering the Alignment | ■ ■ | ■ | Beginner |
| 7 | Creating Groups | ■ ■ | ■ | Beginner |
| 8 | Editing Alignments | ■ ■ | | Beginner |
| 9 | Keyboard Edits | ■ ■ | | Beginner |
| Changing Appearance | | | | |
| 10 | Colouring Alignment | ■ ■ | ■ | Beginner |
| 11 | Altering Alignment Layout | ■ ■ | ■ | Beginner |
| 12 | Graphical Output | ■ ■ | ■ | Beginner |
| 13 | User Defined Colour Schemes | ■ ■ | ■ | Beginner |
| Multiple Sequence Alignment | | | | |
| 14 | Multiple Sequence Alignment | ■ ■ | ■ | Beginner |



Thank you

for your attention