

ASSIGNMENT 2

Dataset 2 ■ Design of a study scheme from available sequence information of influenza viruses and collecting data.

- Phylogeny analyses by MEGA-software (neighbor-joining and UPGMA trees and parsimony cladogram). No mrBayes- analyses with this dataset (you can do them, if you want). If enough time, network-phylogies.

■ Time schedule:

- Wed 2. Feb from 12 to 20 there will be extra help for you in C128. This concerns data collection and alignments for datasets 1 and 2. The week after (maybe in Wed 9. Feb, need to be agreed) you can again come to meet extra help and check your alignments.

- Data-collection and MEGA-analyses ready during weeks (3) and 4.
Work should be ready in 22. Feb.

- Recommendation is that you don't work alone, instead form groups.

ASSIGNMENT 2 – THE FLU-DATABASE

<http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>

Influenza virus resource

HOME SEARCH SITE MAP **Flu home** Database Genome Set Alignment Tree BLAST Annotation FTP Help Contact us

Influenza Virus Resource presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu sequence analysis, annotation and submission to GenBank. In addition, it provides links to other resources that contain flu sequences, publications and general information about flu viruses.

Read more about: [This resource](#) | [Flu database](#) | [Flu sequence submission to GenBank](#) | [NIAID Influenza Sequencing Project](#) | [Influenza virus biology](#)

NCBI

- Growth of flu sequences
- GenBank sequences from the NIAID Project
- Assembly Archive
- Trace Archive
- NIAID data releasing status
- RefSeq genomes
- RefSeq proteins
- Protein Structures
- Flu resources**
- NIAID Project
- JCVI Flu
- Influenza Research Database
- CDC Flu
- Vaccine Selection
- WHO Flu
- NCBI Viruses**
- Viral Genomes
- Virus Variation

Sequences from pandemic (H1N1) 2009 viruses

NEW 12/01/2010: Sequences of WHO recommended vaccine strains can now be selected in the database through "Additional filters". [More news](#)

Search the Flu database, build queries, retrieve sequences, apply analysis tools **Begin**

Database	Select influenza sequences by virus, subtype, host, and other criteria	▶
Genome Set	Find complete genome sets	▶
Alignment	Align your sequence(s) and others in the database (up to 1000 sequences)	▶
Tree	Build a clustering or phylogenetic tree	▶
BLAST	BLAST a flu sequence against the database	▶
Annotation	Annotate flu genomic sequences	▶
Submission	Submit flu sequences to GenBank	▶
FTP	Retrieve database and sequence data through ftp	▶

Sample Searches **Begin**

- Full-length *HA* proteins of the *H3* subtype *Influenza A virus* in *USA* from *1998* to *2002*. ▶
- Complete genome sets of the *Pandemic (H1N1) 2009 Influenza A virus* from *Japan* with the "H274Y" *Drug-resistance mutation* in the NA protein. ▶

New records in Entrez

- [Publications on influenza viruses from the past 2 weeks](#)
- [All influenza sequences updated in GenBank from the past week](#)

ASSIGNMENT 2 – THE FLU-DATABASE

Influenza Virus Sequence Database

Protein or nucleotide sequences can be retrieved from the database using GenBank accession numbers or search terms.

Multiple queries can be built by clicking the "Add Query" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database.

[Permanent link for this query](#)

Get sequences by accession

Upload ## Enter a comma or space separated list of sequence accessions or upload text file with this list.

Select sequence type:

Protein Protein coding region Nucleotide

Search for keyword:

Keyword Search in

Define search set:

Type	Host	Country/Region	Protein	Subtype	Sequences length	Collection date	Release date
any	any	any	any	H any N any	Min.: <input type="text"/> Max.: <input type="text"/>	From: <input type="text"/> <input type="text"/> <input type="text"/> To: <input type="text"/> <input type="text"/> <input type="text"/>	<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>
A	Avian	regions	PB2	1		Year Month Day	Year Month Day
B	Blow fly	Africa	PB1	2			
C	Camel	Asia	PB1-F2	3	<input type="checkbox"/> Full-length only		

- The recent pandemic H1N1 is an A-virus which has type 1 of H (haemagglutinin) gene and type 1 of N (neuraminidase) gene. Within type 1 (and other types as well, for example the bird flu H5N1) very much sequence evolution occurs => lots of information for tracing the virus (its movement, geographical, temporal differences etc.)

ASSIGNMENT 2 – THE FLU-DATABASE

- Construct a comparative plan, for example H1N1 in different countries or continents, maybe at a certain time, temporal changes within a country / continent, virus sequences from human vs animals, or H1N1 vs H5N1, whatever you like, based on the existing information.
- Take as many sequences as you like. Working with more than 500 is not very practical and alignments are very time-consuming.
- It is not a good idea to take many identical or nearly identical sequences (how to avoid that, is your problem.....).
- Restrict your study to H (taking both H and N is not practical, too much work).
- Take only "full-length" sequences (if you include one sequence from which, say, 200 nucleotides are missing from the 3' end, this part is not considered in your phylogeny analyses, i.e. you miss the information from your whole data).
- Probably not much editing after alignments, some sequence items have a short "extra" part at the 5' end. Discard that.

ASSIGNMENT 2 – THE FLU-DATABASE

- In the query result below, from one virus "particle" from Bamako information from many genes: In addition to your study object, H (HA) there are the genes PB2, PA, NP, NA, M1, M2, NS1, NS2. From the samples of Senegal, Djibouti, South Africa and Egypt only HA.

<input checked="" type="checkbox"/>	ADM14980	219	Human	NS1	H1N1	Ethiopia	2009/12/01	Influenza A virus (A/Addis Ababa/WR2848T/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14981	121	Human	NS2	H1N1	Ethiopia	2009/12/01	Influenza A virus (A/Addis Ababa/WR2848T/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14955	759	Human	PB2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14956	716	Human	PA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14957	566	Human	HA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14958	498	Human	NP	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14959	469	Human	NA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14960	252	Human	M1	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14961	97	Human	M2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14962	219	Human	NS1	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ADM14963	121	Human	NS2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
<input checked="" type="checkbox"/>	ACT85993	29	Human	NP	H1N1	Mauritius	2009/06/25	Influenza A virus (A/Candos/1/2009(H1N1))
<input checked="" type="checkbox"/>	ACG50709	466	Human	NA	H1N1	South Africa	2008/05/27	Influenza A virus (A/CapeTown/26/2008(H1N1))
<input checked="" type="checkbox"/>	ACG50719	325	Human	HA	H1N1	South Africa	2008/05/27	Influenza A virus (A/CapeTown/26/2008(H1N1))
<input checked="" type="checkbox"/>	CAD29928	346	Human	HA	H1N1	Senegal	1997	Influenza A virus (A/Dakar/11/97(H1N1))
<input checked="" type="checkbox"/>	CAD29924	355	Human	HA	H1N1	Senegal	2000	Influenza A virus (A/Dakar/17/2000(H1N1))
<input checked="" type="checkbox"/>	ADG21157	566	Human	HA	H1N1	Djibouti	2009/09	Influenza A virus (A/Djibouti/N11092/2009(H1N1))
<input checked="" type="checkbox"/>	ADG21185	566	Human	HA	H1N1	Djibouti	2009/12/08	Influenza A virus (A/Djibouti/N13142/2009(H1N1))
<input checked="" type="checkbox"/>	ABQ53688	470	Human	NA	H1N1	South Africa	1997	Influenza A virus (A/Durbin/113/1997(H1N1))
<input checked="" type="checkbox"/>	CAD29908	363	Human	HA	H1N1	Egypt	2001	Influenza A virus (A/Egypt/101/2001(H1N1))
<input checked="" type="checkbox"/>	ADM26271	566	Human	HA	H1N1	Egypt	2010/01	Influenza A virus (A/Egypt/N00124/2010(H1N1))

ASSIGNMENT 2 – THE FLU-DATABASE

- When you click the accession number of a certain sequence, you first get the protein (amino acid) sequence.
- Note the link to nucleotide sequence and links to many other facilities.

```
AUTHORS   Lin,Y.
TITLE     Direct Submission
JOURNAL   Submitted (19-APR-2002) Lin Y., Virology, National Institute for
          Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UNITED
          KINGDOM

FEATURES  Location/Qualifiers
  source   1..363
           /organism="Influenza A virus (A/Egypt/101/2001(H1N1))"
           /strain="(A/Egypt/101/2001(H1N1))"
           /serotype="A"
           /db_xref="taxon:192544"
  Protein 1..363
           /product="haemagglutinin"
  Region 20..>363
           /region_name="Hemagglutinin"
           /note="Hemagglutinin; pfam00509"
           /db_xref="CDD:109560"
  CDS    1..363
           /gene="HA"
           /coded_by="AJ457871.1:<4..>1092"
           /db_xref="GOA:Q8AZG4"
           /translation="MSEK..."
```

All links from this record

- [BLink](#)
- [Related Sequences](#)
- [CDD Search Results](#)
- [Conserved Domains \(Concise\)](#)
- [Conserved Domains \(Full\)](#)
- [Domain Relatives](#)
- [Encoding mRNA](#)
- [Nucleotide](#)
- [PubMed](#)
- [Related Structures \(List\)](#)
- [Related Structures \(Summary\)](#)
- [Taxonomy](#)

Nucleotide records that are the source of protein records in the current set. The protein sequences are generated through translation of coding region features on the nucleotide records.

- FASTA-link is at the left up corner of the page.

ASSIGNMENT 2 – THE FLU-DATABASE

- You get the sequence in FASTA-format like this:

```
>gi|22859141|emb|AJ457871.1| Influenza A virus (A/Egypt/101/2001 (H1N1)) partial  
HA gene for haemagglutinin, genomic RNA  
AAAATGAAAGCAAACTACTGGTCCTGTTATGTACATTTACAGCTACATATGCAGACACAATATGTATAG  
GCTACCATGCAAACAACCTCAACCGACACTGTTGACACAGTACTTGAGAAGAATGTGACAGTGACACACTC  
TGTC AACCTACTTGAGGACAGTCACAATGGAAAACCTATGTCTACTAAAAGGAATAGCCCCACTACAATTG  
GGTAATTGCAGCGTTGCCGGATGGATCTTAGGAAACCCAGAATGCGAATTACTGATTTCCAAGGAATCAT  
GGTCCTACATTGTAGAAAACCAAATCCTGAGAATGGAACATGTTACCCAGGGTATTTGCGCCGACTATGA  
GGA ACTGAGGGGAGCAATTGAGTT CAGTATCTTCATTTGAGAGGTTTGAAAATATTTCCCAAGAAAAGCTCA  
TGGCCCAACCCACCCGTAACCGGAGTATCAGCATCATGCTCCCATAATGGGAAAAGCAGTTTTTACAGAA  
ATTTGCTATGGCTGACGGGGAAGAATGGTTTGTACCCAAACCTGAGCAAGTCTTATGCAAACAACAAAAGA  
GAAAGAAGTCTTTGACTATGGGGTGTTCATCACCCGCCTAACATAGGGGACCAAGGGCCCTCTATCAT  
ACAGAAAATGCTTATGTCTCTGTAGTGTCTTACATTATAGCAGAAGATTACCCCCAGAAAATAGCCAAAA  
GACCCAAAAGTAAGAGGT CAGGAAGGAAGAATCAACTACTACTGGACTCTGCTGGAACCCGGGGATACAAT  
AATATTTGAGGCAAATGGAAATCTAATAGCGCCAAGGTATGCTTTGCACTGAGTAGAGGCTTTGGATCA  
GGAATCATCACCTCAAATGCACCAATGGATGAATGTGATGCGAAGTGTCAAACACCTCAGGGAGCTATAA  
ACAGCAGTCTTCTTTCCAGAATGTACCCAGTCACAATAGGAGAGTGTCCAAAGTATGT CAGGAGTGC  
AAAATTAAGGATGGTTACAGGACTAAGGAACATCCCATCCATTCAATCCAGAGGTTTTGTTTGGAGCCATT  
GCCGGTTTTATTGAAGGGGGGTGGACTGGAATGATGGATGGG
```

- The title (what follows the mark >) is too long and in your subsequent analyses (already after Clustal - alignment) you see only some first part of it .
- Edit the title into a reasonable form, discard everything that is not necessary!

- This is an example of a reasonable title: >Egypt/101/2001 AJ457871