Assignment 2 / Biometry and bioinformatics I / 2013

Monitoring/tracing of viruses and bacteria on the basis of their informative sequencies is done very actively, providing understanding of their behaviour, for example spread of an epidemia.

In this assignment you get familiar with one database (out of many different kind of virus- and bacteria sequence databases) and use it´s information (seqs) for studying the behavior of the influenza-virus H1N1 which caused a pandemia in 2009[*].

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

The study question: on the basis of existing sequence information, what can be inferred about the behaviour of H1N1 by sequence clustering
- before the pandemia -> 2008
- during the pandemia 2009 and after that (2010 ->)

Part (a)
First you perform a small experiment by collecting a small set of seqs from 2-3 countries, from different time points. Small set = 1-2 seqs as representatives of one country and one time point.
Align the seqs and cluster them by UPGMA.
You get some kind of an idea to perform a more comprehensive study.
In the report give this UPGMA-tree and your interpretation.
For example “it seems to be that clustering (sequence similarities – differences) is based on (x) geographical location , (y) time; you very probably notice that either (x) or (y) is the predominant “apparent explanation”. Describe your explanation (i.e. “the seems to be” –situation)

Part (b)
On the basis of your “seems to be” –experiment (part (a) you make a plan for a “real study” (part (b)) by collecting more data.
The amount of data is your own choice. Maybe more seqs (now several tens; in part (a) you had one or two) from those countires which you included in your experiment, maybe other countries (nearby vs. remote), maybe more time points.
Then you again align the seqs, perform UPGMA clustering and write you interpretation about the epidemic behavior of H1N1.

Your report should thus include UPGMA-tree with explanations from part (a) and UPGMA-tree with explanations from part (b). And the study plan which you made, on the basis of part (a) to do part (b). A study plan very probably includes modifications to your original plan because of lack of data.

[*] Extensive sequence information, though not complete (in fact, far of complete): if a database has, say, 100 sequencies from a given virus from a given country from a given time point (year, for example), this does not mean that there were 100 infections. It means that 100 sequences have determined, according to some criteria, according to some interests within the scope of research financiation.

Practical advise

H1N1 is one of many influenza A-virus subtypes.
Influenza A-virus in composed of 11 genes, H (HA) and N (NA) being those genes which serve as defining a given subtype.
H = hemaglutinin gene (HA)
N = neuraminidase gene (NA)

When a virus is typed by sequencing, usually at least the H-gene is sequenced => most information is from the H-gene.

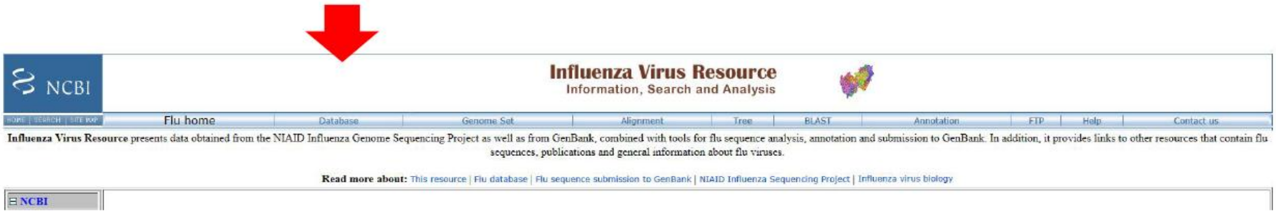
We restrict this course assignment to H-gene sequencies from H1N1 and we work with DNA-information (nucleotide information), not protein (amino acid information).

Inspect carefully the contents of next page: how to find H-gene nucleotide seqs, i.e. what boxes you should click in the database in order to define that you want to get H-seqs (you must click HA-protein) from the H1N1 type of A-influenza virus (you must click 1 from the H-subtype box and 1 from the N-subtype box). And, when collecting data, you should download "nucleotide", not "protein".

Practical advice for collecting data from Flu-database

http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
The main page has links to all kind of updated information about flu-viruses

Go here

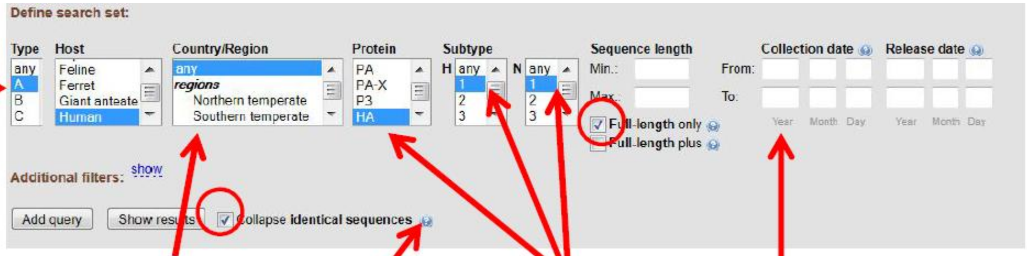


Influenza Virus Resource presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu sequence analysis, annotation and submission to GenBank. In addition, it provides links to other resources that contain flu sequences, publications and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

Read more about: This resource | Flu database | Flu sequence submission to GenBank | NIAID Influenza Sequencing Project | Influenza virus biology

(The database has many facilities. Such as aligning data, drawing phylogenetic trees. You can, of course, try them. However, they do not operate very well.....)

Influenza A-virus



Define search set:

Type	Host	Country/Region	Protein	Subtype H	Subtype N
any	Feline	any	PA	any	any
A	Ferret	*regions*	PA-X	1	1
B	Giant anteate	Northern temperate	P3	2	2
C	Human	Southern temperate	HA	3	3

Sequence length Min.: Max.:

☑ Full-length only

□ Full-length plus

Collection date From: To: Year Month Day

Release date Year Month Day

Additional filters: show

Add query | Show results | ☑ Collapse identical sequences

Select from here countries/regions you want to study.

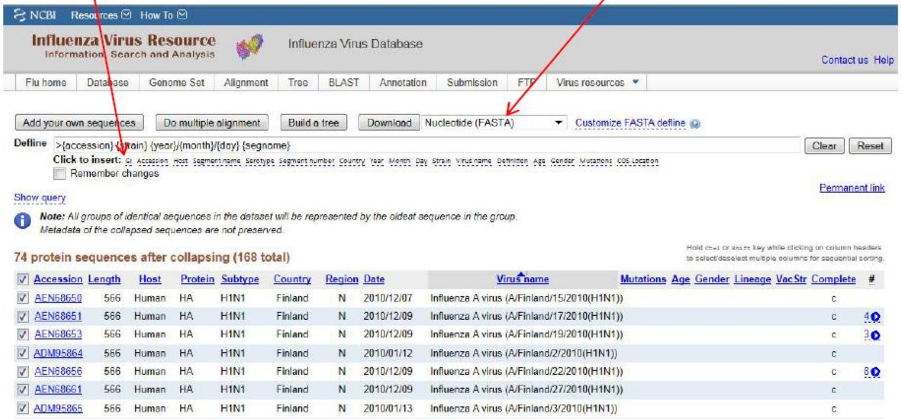
Read this . Not a good idea to collect many identical copies. On the other hand, which seqs are present as many copies, is valuable information.

These three clicked on specify that you will get hemaglutinin seqs from the type H1N1-virus

2009-2013. Lots of data from 2009, some data may exist also 2010 ->

An example, results from Finland, you need nucleotide data, click from here
From here you can define the information appearing in the title of each seq

It is not a good idea to keep the title (= text appearing after >... too long). Include country, year, month, accession number.



NCBI Resources How To
Influenza Virus Resource
Information, Search and Analysis
Influenza Virus Database
Contact us Help
Flu home
Database
Genome Set
Alignment
Tree
BLAST
Annotation
Submission
FTP
Virus resources
Add your own sequences
Do multiple alignment
Build a tree
Download
Customize FASTA defline
Clear
Reset
Click to insert: GI Accession Host Segment name Serotype Segment number Country Year Month Day Strain Virus name Definition Age Gender Mutations CDS Location
Permanent link
Show query
Hold ctrl or shift key while clicking on column headers to select/deselect multiple columns for sequential sorting.

Nucleotide (FASTA)

Defline >{accession} {strain} {year}/{month}/{day} {segname}

□ Remember changes

Note: All groups of identical sequences in the dataset will be represented by the oldest sequence in the group. Metadata of the collapsed sequences are not preserved.

74 protein sequences after collapsing (168 total)

☑	Accession	Length	Host	Protein	Subtype	Country	Region	Date	Virus name	Mutations	Age	Gender	Lineage	VacStr	Complete	#
☑	AEN68650	566	Human	HA	H1N1	Finland	N	2010/12/07	Influenza A virus (A/Finland/15/2010(H1N1))						c	
☑	AEN68651	566	Human	HA	H1N1	Finland	N	2010/12/09	Influenza A virus (A/Finland/17/2010(H1N1))						c	4
☑	AEN68653	566	Human	HA	H1N1	Finland	N	2010/12/09	Influenza A virus (A/Finland/19/2010(H1N1))						c	3
☑	ADM95864	566	Human	HA	H1N1	Finland	N	2010/01/12	Influenza A virus (A/Finland/2/2010(H1N1))						c	
☑	AEN68656	566	Human	HA	H1N1	Finland	N	2010/12/09	Influenza A virus (A/Finland/22/2010(H1N1))						c	8
☑	AEN68661	566	Human	HA	H1N1	Finland	N	2010/12/09	Influenza A virus (A/Finland/27/2010(H1N1))						c	
☑	ADM95865	566	Human	HA	H1N1	Finland	N	2010/01/13	Influenza A virus (A/Finland/3/2010(H1N1))						c	

Aligning seqs: You can use either Clustal or the MAFFT-server. In both cases you take the FASTA-file, now aligned, from the result-link and start operating with it. First you must delete the extra / unnecessay and confusing part from the beginning (if such exists): You can easily see that while most seqs start from the *real start position of the gene*, ATG..., some seqs have something extra before ATG. Check how many nucleotides. This means that you pick up the file to MEGA or Clustal (or some other program by which you can see it and identify the start position of the gene ATG..., most seqs start from this, but not all): remove the extra block and save it to a new FASTA-file and work further with that file.

Your report should also include your aligned FASTA-files, like in assignment 1.