# Assignment 2 / Biometry and bioinformatics I / 2014

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 epidemy. In assignment set 4 you will get familiar with various kind of examples by reading and writing short essay(s) or comments about scientific publications.

In this assignment 2 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.

- The database includes extensive sequence information. What does this mean? If a database has, say, 100 seqs from the 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, for example. One seq (with an accession number, is a virus from one infected human (or animal) individual.

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 -> )
- The most active period of the pandemia was Autumn 2009. Your task is to first pick up countries which you would like to study, and from which information exist, perform an initial study (part (a)) and then a more comprehensive study (part (b)).
- The pandemia started from Mexico in Spring 2009. There is lots of data from various countries from 2009 and also 2010 ->. But very little data from the period before 2009. You don´t have to restrict your study so that all your countries should have pre-pandemic data! Take at least one country from which there is pre-pandemic data, and data from 2009 and 2010 ->. And then additional countries with data from 2009 and 2010 ->.

- Demo of the flu-database during the first course session (9.9) => starting data collection.
- At the beginning of the second course session a short demo about MEGA5, the software for sequence clustering => clustering by neighbor joining algorithm can be done after this demo; a more detailed lecture about this kind of sequence processing methods will be in session 23.9.

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 the neighbor joining method.

- You get some kind of an idea to perform a more comprehensive study.
- In your final report give also this initial neighbor joining tree and explain your initial 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 some 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 neighbor joining clustering and write you interpretation about the epidemic behavior of H1N1.

Your report should thus include the neighbor joining -tree with explanations from part (a) and neighbor joining -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.

---

Practical advise

H1N1 is one of many different 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.

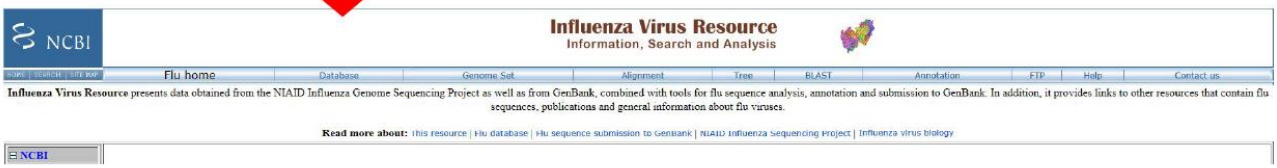
We restrict this course assignment to H-gene sequencies from H1N1 and we work with DNA-information (nucleotide information), not protein (amino acid information).

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".

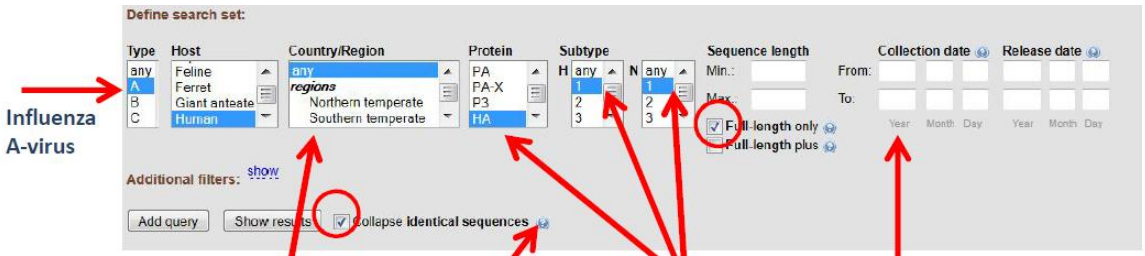
## Practical advice for collecting data from Flu-database

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




(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.....)



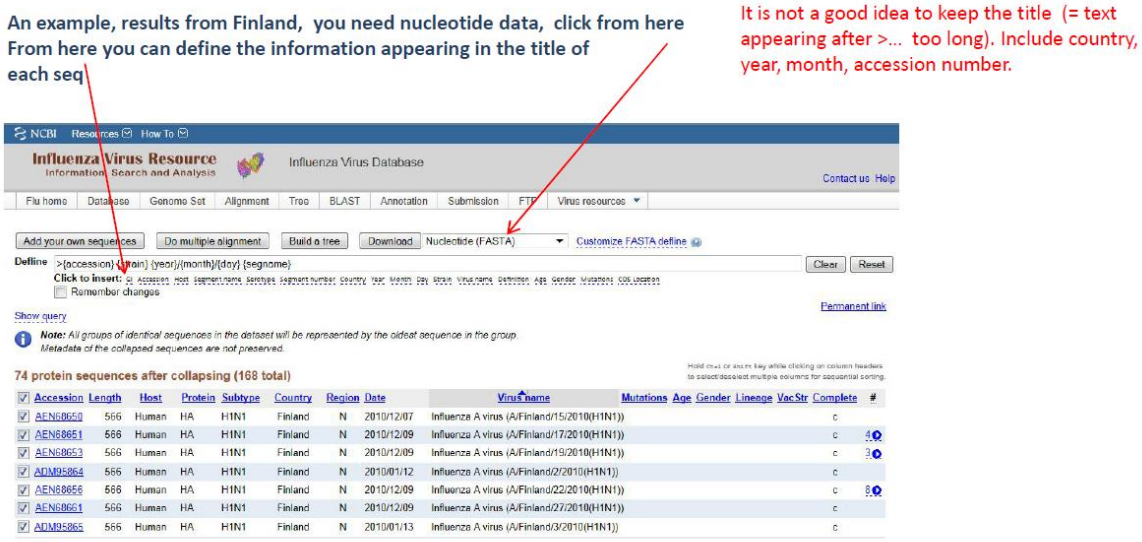

Influenza A-virus

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 ->




Aligning seqs:

- You can use either Clustal (installed in classroom C128 computers; freely downloadable to your own computer, too) 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 / unnecessary 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 (in those cases where some extra nucleotides exist, there are usually up to 46 bp extra, check !). Remove the extra block and save the edited (= extra nucleotides removed) file as a new FASTA-file and work further with that file.

Report from Assignment 2 to be submitted to course Moodle during the first week of October. Your report should also include your aligned FASTA-files as separate text documents. In your report document: explain what you done and your interpretations about the behaviour of H1H1 on the basis of your study scheme. Copy-paste your neighbor joining clustering tree(s) into your report (DO NOT include them as separate MEGA5-files.)

During the last course session we shall see look together your work and also in the light of scientific literature about H1H1.

One of the extra assignments, for extra credits (i.e. extending the 5 cr -> more credits to added later without a deadline) will a more comprehensive study on H1N1 also including scientific literature as a baseline.