Biodatabases / exercises Tuesday 20.1.09

Some comments on the basis of the answers I have received:

´Picking up´ defined sequences from the flu database was apparently very easy and all answers were correct. However, the question ´give references to those publications….´ was not clear…. My question was (perhaps) not formulated well enough: "publication" means a scientific article which has authors (the scientists who have performed the research, which includes sequence information which they have deposited to genbank database), name of the article and name of the scientific journal which has published the work. Pubmed is the database which takes care of biomedical scientific publications. (So, maybe I should have asked "publications in Pubmed"….). The idea of the question was to check if you know how to go from sequence database (on the basis of accession numbers) to publication database.

This is an example of a correct answer, a scientific publication in J. Gen. Virology:

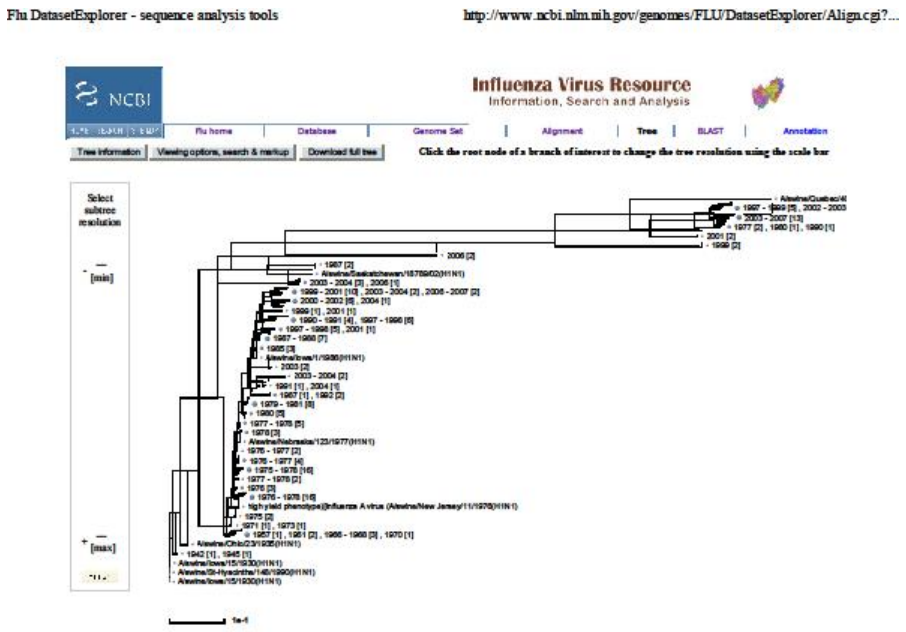
```
LOCUS       AAA64229                566 aa            linear   VRL 03-MAY-
2006
DEFINITION  hemagglutinin [Influenza A virus (A/seal/MA/3911/1992(H3N3))].
ACCESSION   AAA64229
VERSION     AAA64229.1  GI:608638
DBSOURCE    locus FLAHASEALA accession L31949.1
KEYWORDS    .
SOURCE      Influenza A virus (A/seal/MA/3911/1992(H3N3))
  ORGANISM  Influenza A virus (A/seal/MA/3911/1992(H3N3))
            Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
            Influenzavirus A.
REFERENCE   1  (residues 1 to 566)
  AUTHORS   Callan,R.J., Early,G., Kida,H. and Hinshaw,V.S.
  TITLE     The appearance of H3 influenza viruses in seals
  JOURNAL   J. Gen. Virol. 76 (Pt 1), 199-203 (1995)
   PUBMED   7844533
```

This is an example of sequence information which has not been published, it only exists in the sequence database: (Published sequence data and unpublished sequence data are fundamentally different. For example, published sequences are freely available to other scientist, but unpublished are not.)

```
LOCUS       BAF36959                152 aa            linear   VRL 17-NOV-
2006
DEFINITION  polymerase acidic protein [Influenza A virus
            (A/seal/Massachusetts/1/80(H7N7))].
ACCESSION   BAF36959
VERSION     BAF36959.1  GI:118123406
DBSOURCE    accession AB282884.1
KEYWORDS    .
SOURCE      Influenza A virus (A/seal/Massachusetts/1/80(H7N7))
  ORGANISM  Influenza A virus (A/seal/Massachusetts/1/80(H7N7))
            Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
```

```
            Influenzavirus A.
REFERENCE   1
  AUTHORS   Kida,H. and Sakoda,Y.
  TITLE     Genetic and antigenic analyses of H5 influenza viruses from
aquatic
            birds for vaccine and diagnostic use
  JOURNAL   Published Only in Database (2006)
REFERENCE   2  (residues 1 to 152)
  AUTHORS   Tanaka,Y., Kida,H. and Sakoda,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-NOV-2006) Yukiko Tanaka, Hokkaido University,
            Graduate School of Veterinary Medicine; kita-ku,kita18 nishi9,
            Sapporo, Hokkaido 060-0818, Japan
            (E-mail:influ@vetmed.hokudai.ac.jp, Tel:81-11-706-5209,
```

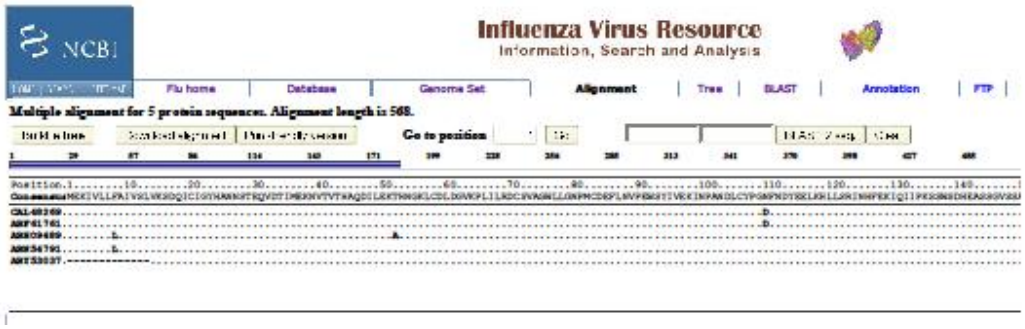
The ´build a tree´ operation should have resulted in the following tree:



If you did not get a tree, you perhaps did not notice that you should have clicked "next step"……. Or you did not follow the piece of advice: "In the exercises restrict the queries to these segments (HA or NA)", and you tried with too many sequences (the limit for this opearation is 1000).

The alignment question should have resulted in the following:

Flu DatasetExplorer - sequence analysis tools http://www.ncbi.nlm.nih.gov/genomes/FLU/DatasetExplorer/JustAlign.cgi



1 2 3

This is the view of the first 150 amino acids (the total length is 560 amino acids). Here you can see amino acid differences at 3 sites (marked below the alignment). Altogether there are 11 variable amino acid sites. (the last sequence has ---------- at the beginning; this means that there is deletion, or (more probably) this particular sequence has not been sequenced from the very beginning (=information lacking).

The purpose of this exercise was to make clear that you understand what is an alignment and how to read it. Comparing sequences = the first step is usually to align them.