Sequence Alignments and Phylogeny

Genomics and Clinical Virology

Sunando Roy

sunando.roy@ucl.ac.uk

Scenario: There has been a recent spike in Influenza cases in your local hospital. The clinicians are worried there may be an outbreak and ward to ward transmission. They have asked you to sequence the virus from clinical samples to determine if they are related.

Outline

- Sequence retrieval from GenBank
- Multiple Sequence Alignments
- Model Testing and Maximum Likelihood based tree building
- Viewing and modifying a Tree file

Software

Mafft – Alignment tool

MEGA – Alignment viewer and editor

Modeltest-ng – Model testing

IQ-TREE – Tree building

Figtree – Tree viewer and editor

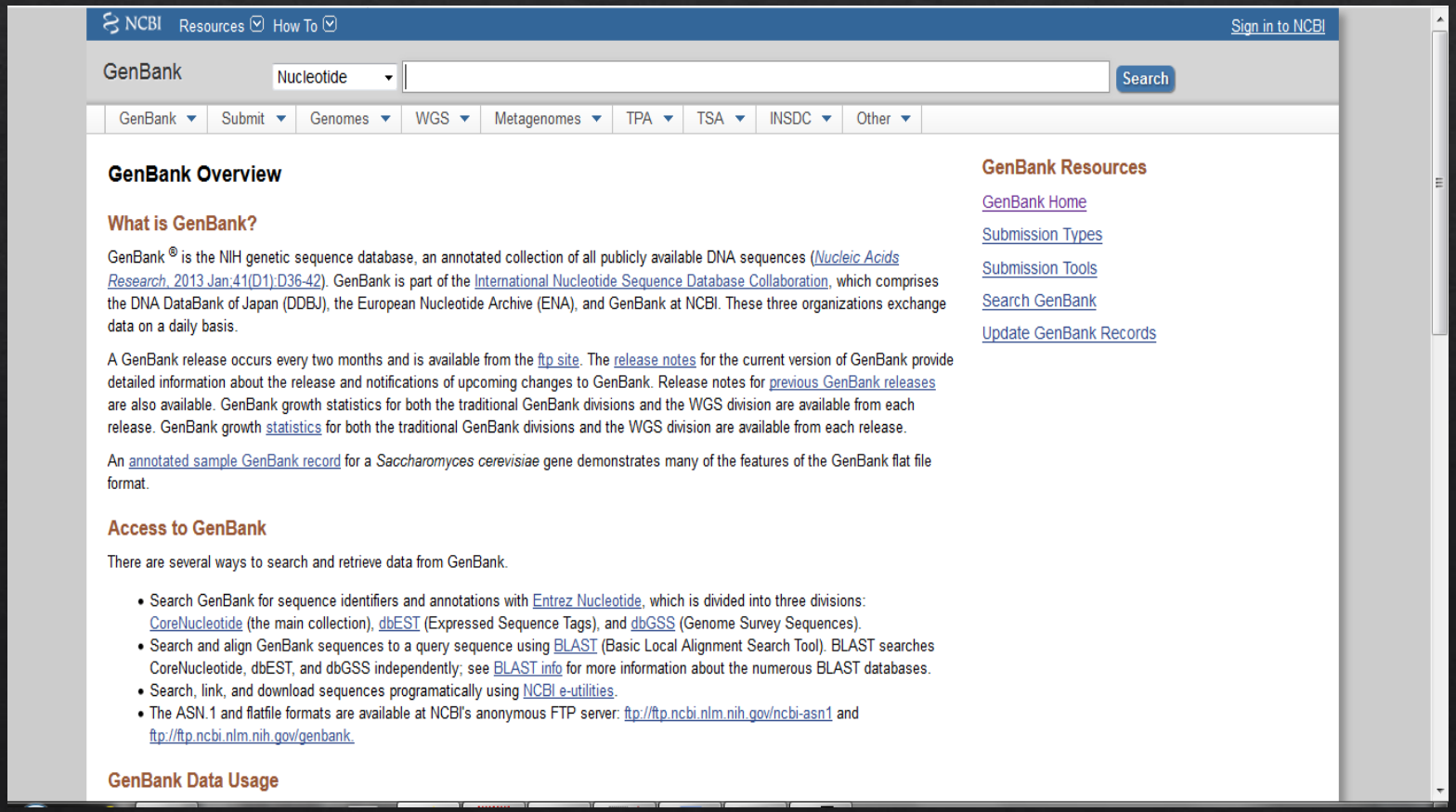
Sequence Retrieval

- **Where?**
  - GenBank (https://www.ncbi.nlm.nih.gov/genbank/)
  - ENA (https://www.ebi.ac.uk/ena)
  - DDBJ (http://www.ddbj.nig.ac.jp/)

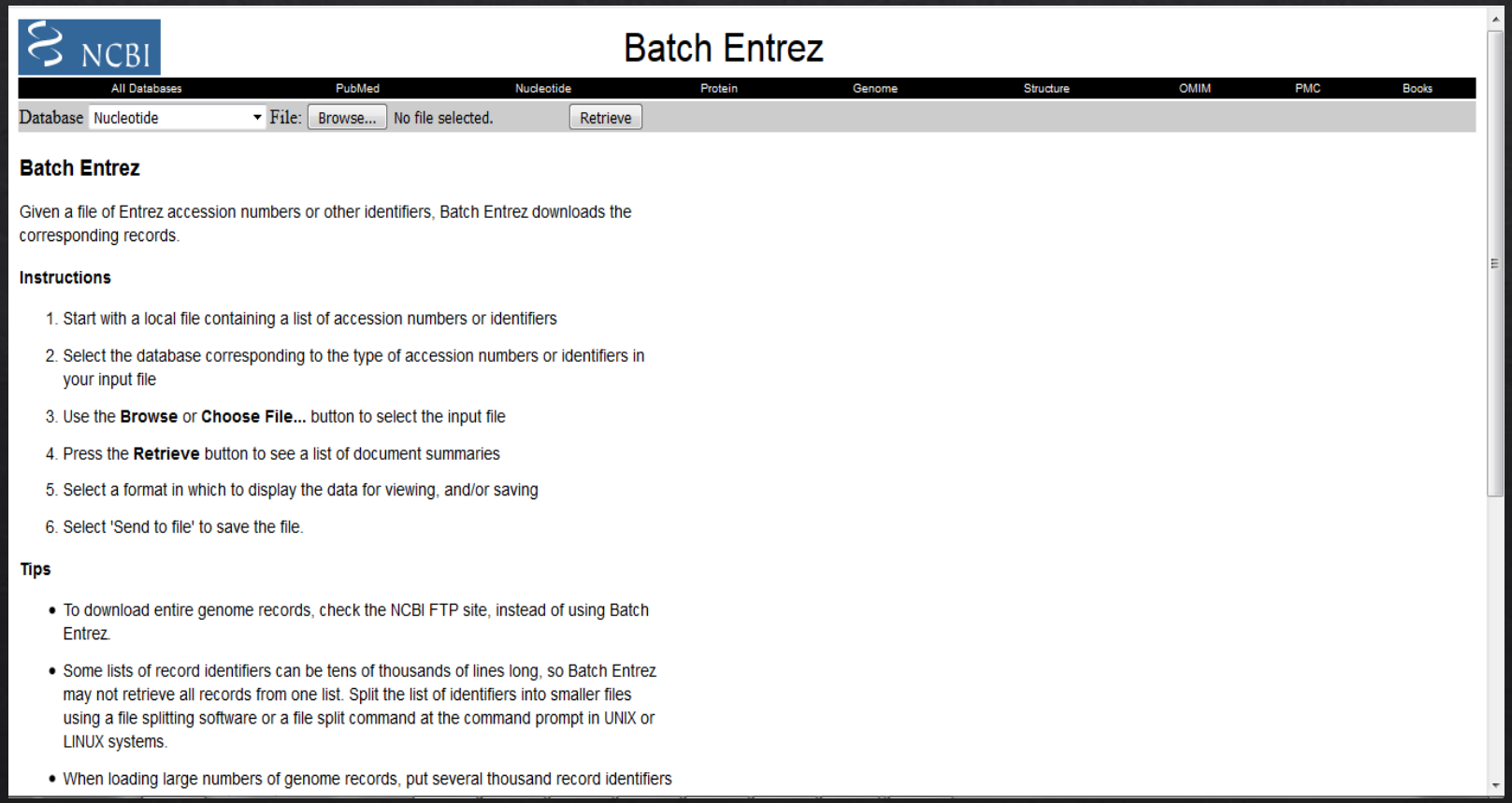
For all other databases
(https://en.wikipedia.org/wiki/List_of_biological_databases)

There are now pathogen specific databases like GISAID (Influenza/SARS-CoV-2) and NoroNet (Norovirus)

- **How?**
  - GenBank (https://www.ncbi.nlm.nih.gov/genbank/)



- **Alternatives**
  - Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez)



- **Alternatives**
  - Entrez E-utilities (ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/)
  - Command – Browser example

    https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=AY278488, AY304486, MN908947, MT782115&rettype=fasta&retmode=text
  - Command – Terminal example

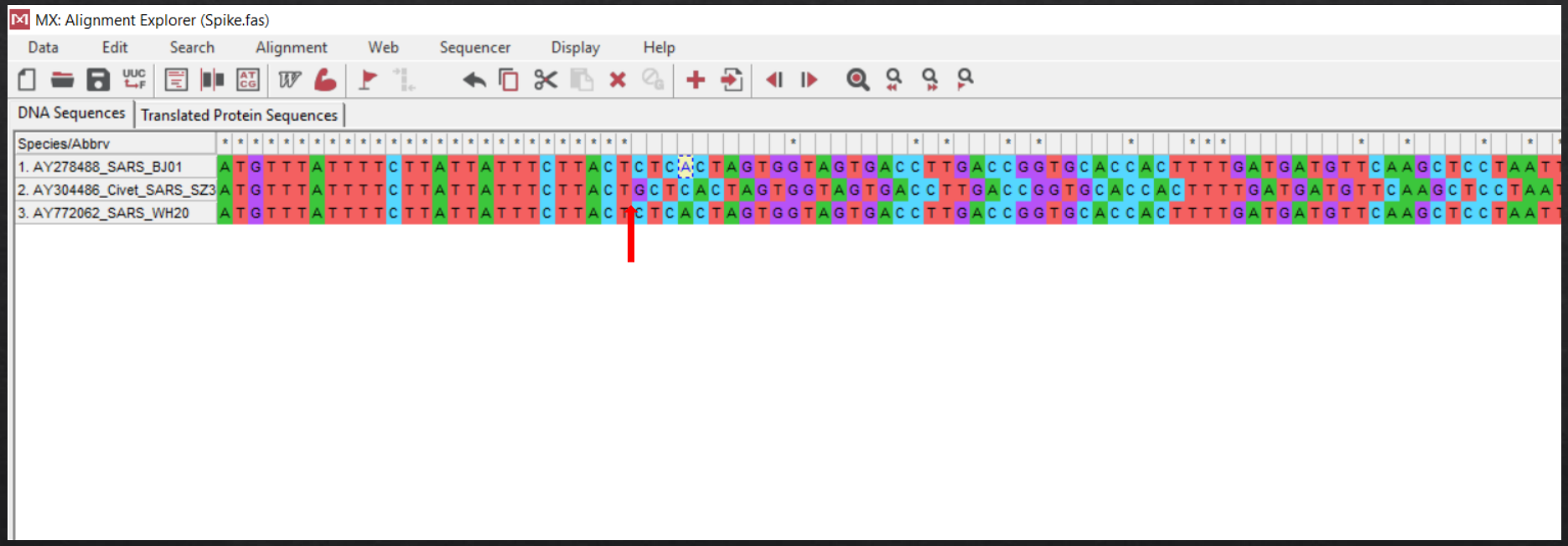
    esearch -db "protein" –query "txid11270[Organism] AND L Protein Complete AND refseq[filter]" | efetch –format fasta > outputfile.fasta

Dataset Selection

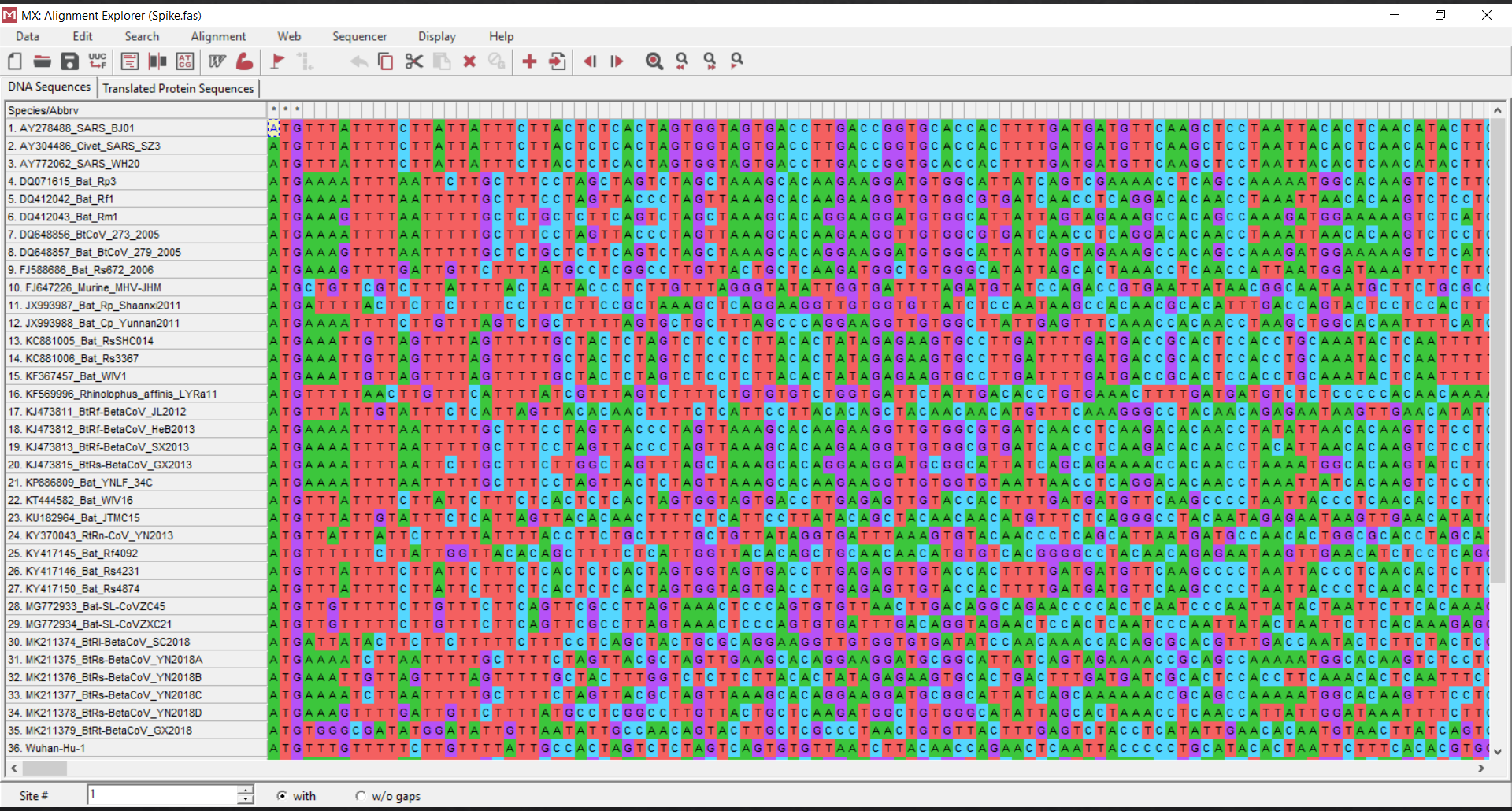
- Relatedness with query strain.
- Time/Location based filtering of strains.
- Size/Completeness of dataset to process.
- Choosing the correct outgroup.

Multiple Sequence Alignments

- **Why?**
  - Necessary for every sequence analysis

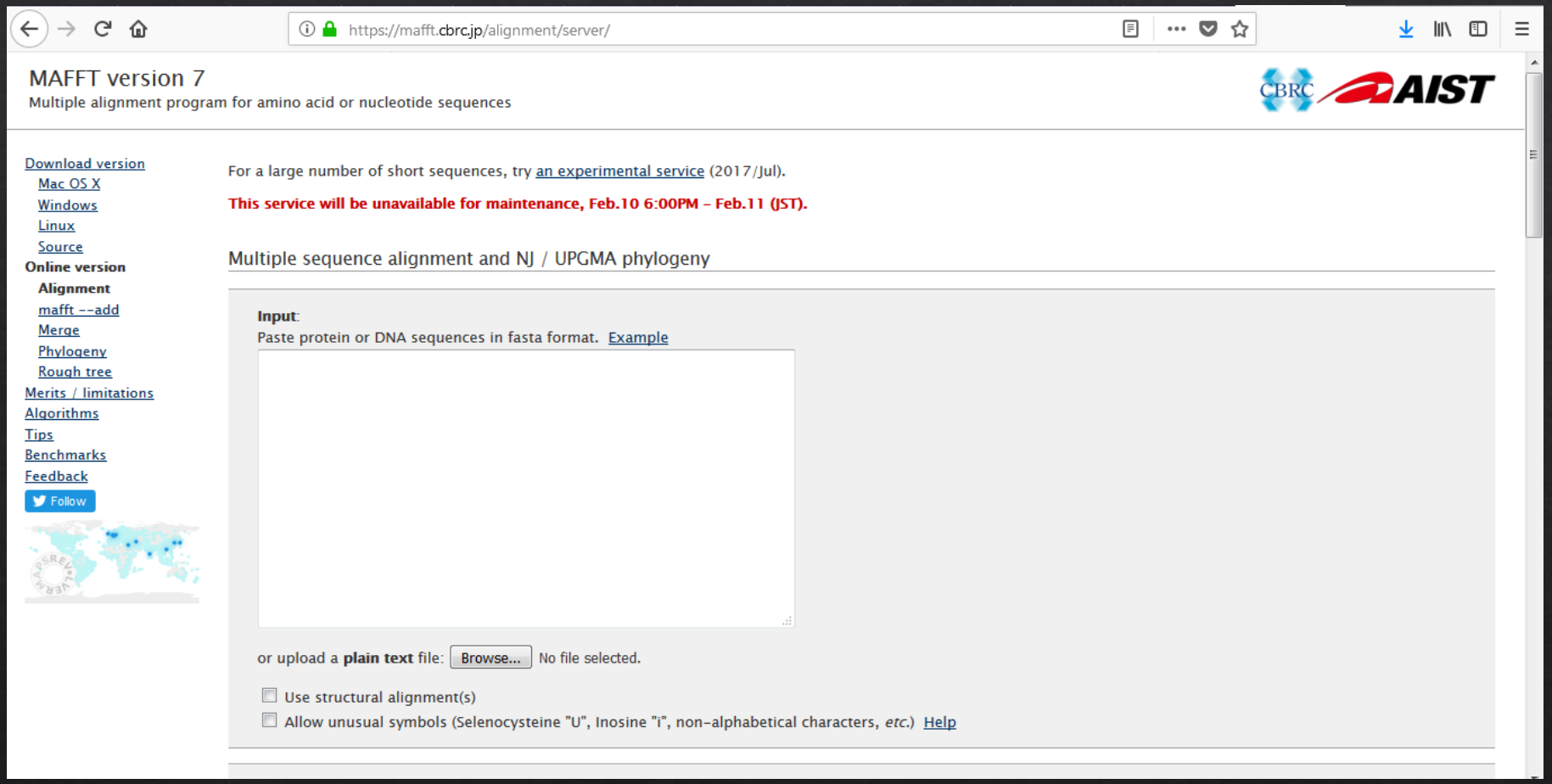



MX: Alignment Explorer (Spike.fas)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences
Species/Abbrv
1. AY278488_SARS_BJ01
2. AY304486_Civet_SARS_SZ3
3. AY772062_SARS_WH20


MX: Alignment Explorer (Spike.fas)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences
Species/Abbrv
1. AY278488_SARS_BJ01
2. AY304486_Civet_SARS_SZ3
3. AY772062_SARS_WH20
4. DQ071615_Bat_Rp3
5. DQ412042_Bat_Rf1
6. DQ412043_Bat_Rm1
7. DQ648856_BtCoV_273_2005
8. DQ648857_Bat_BtCoV_279_2005
9. FJ588686_Bat_Rs672_2006
10. FJ647226_Murine_MHV-JHM
11. JX993987_Bat_Rp_Shaanxi2011
12. JX993988_Bat_Cp_Yunnan2011
13. KC881005_Bat_RsSHC014
14. KC881006_Bat_Rs3367
15. KF367457_Bat_WIV1
16. KF569996_Rhinolophus_affinis_LYRa11
17. KJ473811_BtRf-BetaCoV_JL2012
18. KJ473812_BtRf-BetaCoV_HeB2013
19. KJ473813_BtRf-BetaCoV_SX2013
20. KJ473815_BtRs-BetaCoV_GX2013
21. KP886809_Bat_YNLF_34C
22. KT444582_Bat_WIV16
23. KU182964_Bat_JTMC15
24. KY370043_RtRn-CoV_YN2013
25. KY417145_Bat_Rf4092
26. KY417146_Bat_Rs4231
27. KY417150_Bat_Rs4874
28. MG772933_Bat-SL-CoVZC45
29. MG772934_Bat-SL-CoVZXC21
30. MK211374_BtRl-BetaCoV_SC2018
31. MK211375_BtRs-BetaCoV_YN2018A
32. MK211376_BtRs-BetaCoV_YN2018B
33. MK211377_BtRs-BetaCoV_YN2018C
34. MK211378_BtRs-BetaCoV_YN2018D
35. MK211379_BtRt-BetaCoV_GX2018
36. Wuhan-Hu-1
Site #
1
with
w/o gaps

- **How?**
  - Maximize identity between sequences in your alignment.
  - Uses scores for matches (+), mismatches (-), gap penalty (-).
  - Changing scoring parameters change alignments.
  - Visual inspection is always necessary.
  - Bad alignment = Bad phylogenetic inferences.

https://mafft.cbrc.jp/alignment/server/




```
mafft Combined_HA_genbank.fas > outputfile.fas
```

Model Testing and Maximum Likelihood based tree building

- **Why?**
  - To infer evolutionary relationships and identify novel pathogens.
  - Identify geographical location for the source of infection.
  - Identify potential host species.
  - Infer transmission events and outbreaks.

Tree Building

Seq 1 – ATTGCAAT

Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

Tree Building

Seq 1 – ATTGCAAT

Seq 2 – ATTGCAAT

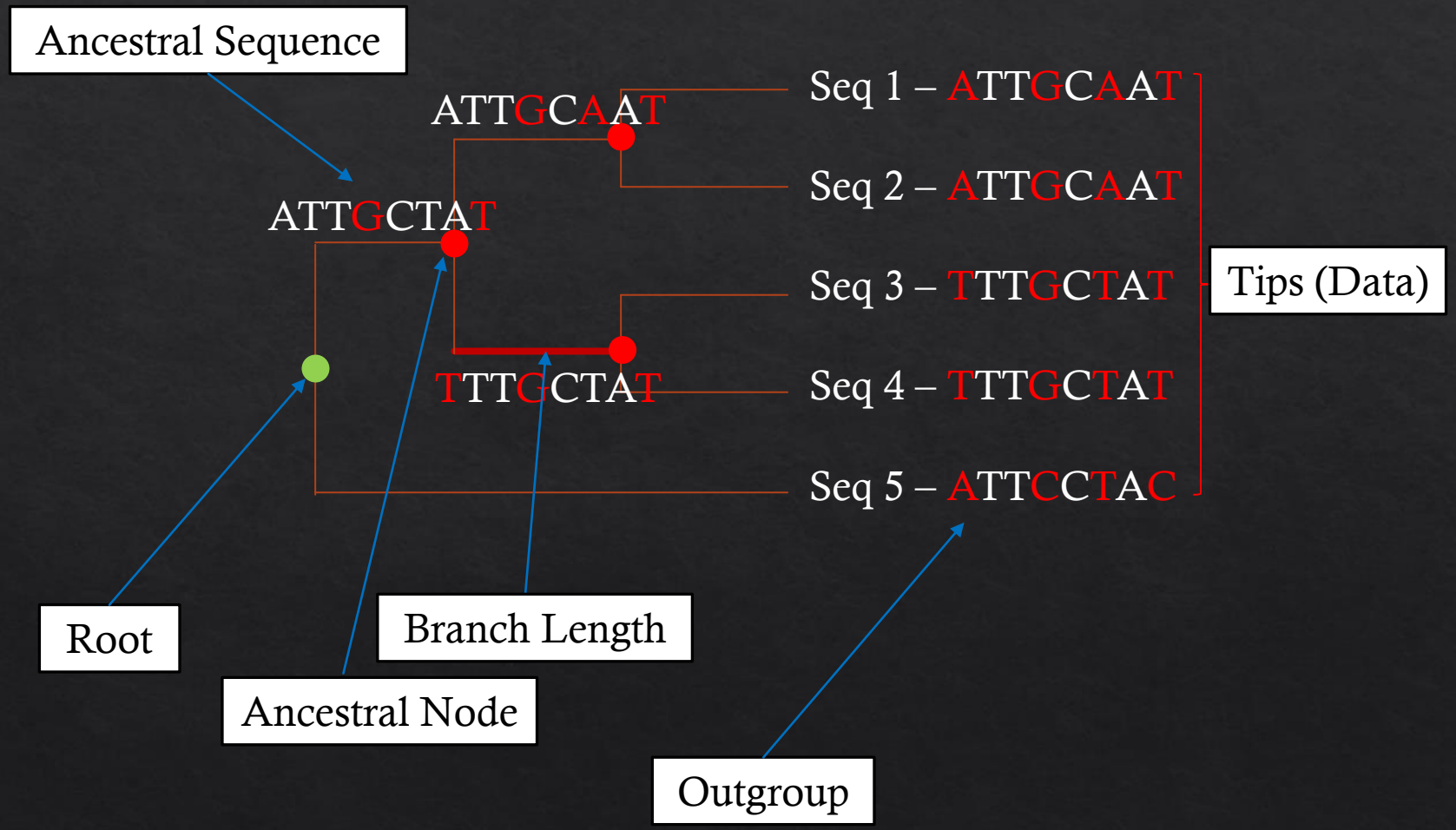
Seq 3 – TTTGCTAT

Seq 4 – TTTGCTAT

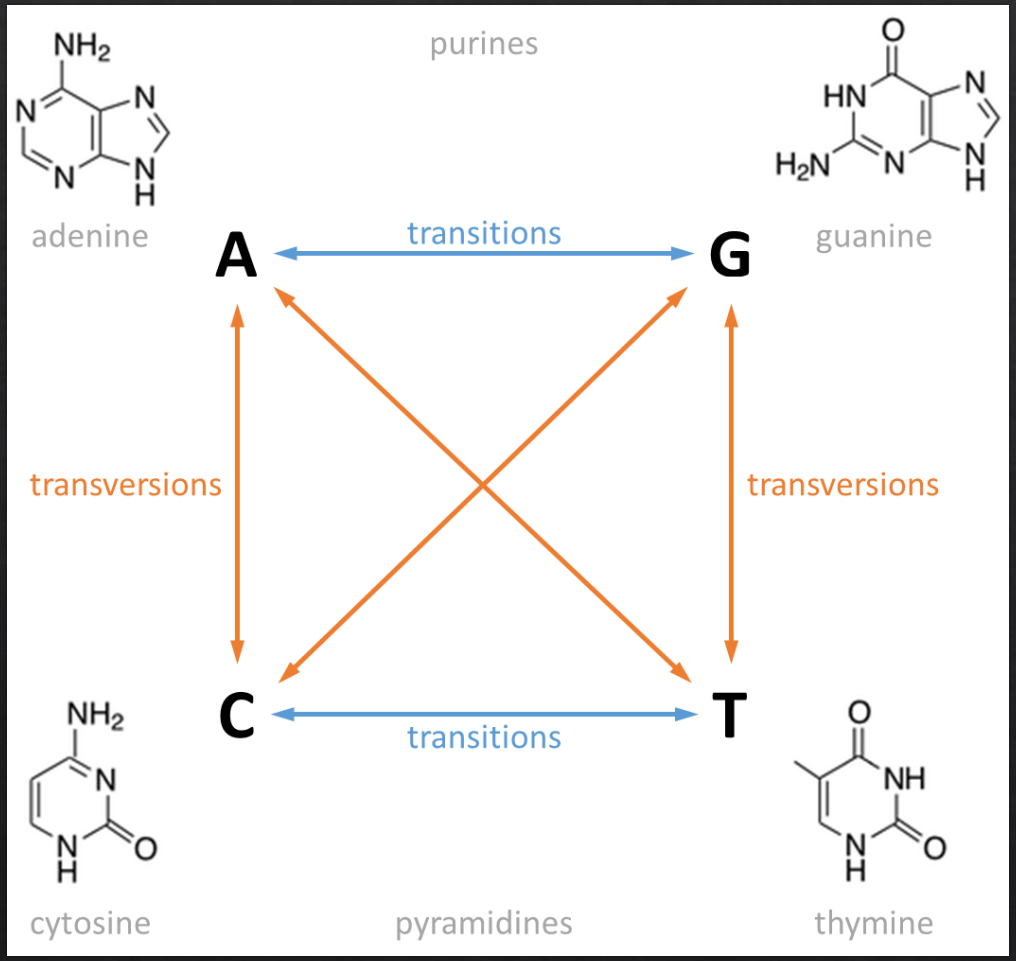
Seq 5 – ATTCCTAC

Tree Building

Tree Building



Evolutionary models - DNA



Jukes Cantor

Complexity

General Time Reversible

Evolutionary models - Proteins

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

https://en.wikipedia.org/wiki/BLOSUM

- Tree Building Tools
  - Neighbour Joining – BioNJ, MEGA
  - Maximum Likelihood – PhyML, RAxML, IQTREE
  - Bayesian Inference – Mr Bayes, BEAST

Neighbour Joining

- Estimates relationships based on a pairwise distance matrix.
- Distance matrix calculation does take into consideration evolutionary substitution models.
- Collapses closest distance pair into one taxa and repeats steps until all tips are clustered.
- Samples only one possible tree out of all possible outcomes.
- Fast but struggles in estimating relationships over longer evolutionary times.

Maximum Likelihood

- Calculates the probability of a tree topology at each individual site across the sequence and a final product across sites is computed.
- Can use independent rate measurement across each site.
- Evaluates multiple tree topologies using branch swaps, nearest neighbour interchange etc. to find trees with the best probability.
- The probability is presented as a log likelihood thus less negative the number the greater the probability.

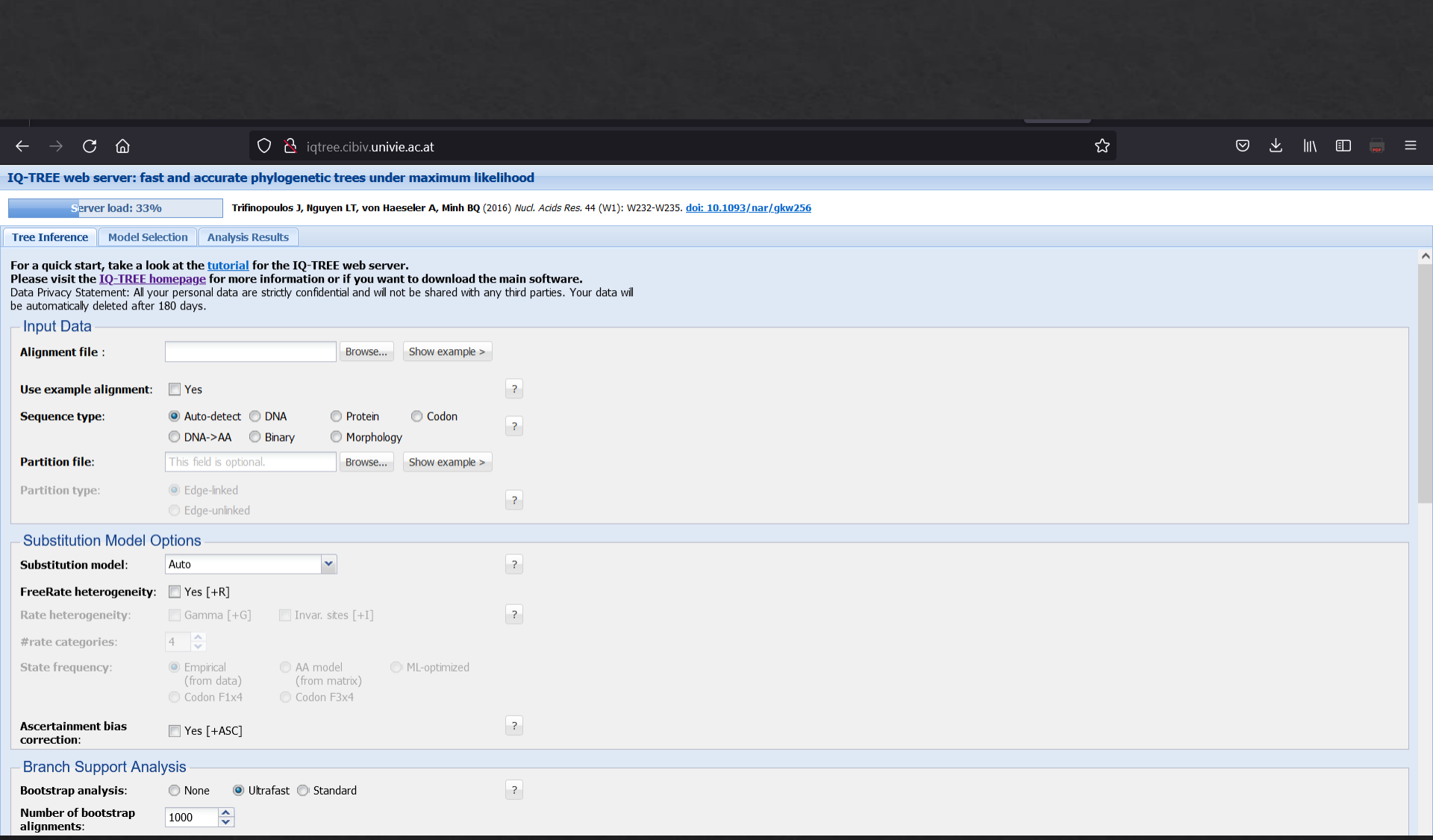
Bayesian Inference

- Trees built by estimating posterior probability from a set of user defined priors.
- Used in Phylodynamics and Phylogeography analysis.
- Computationally intensive.
- Sensitive to priors and evolutionary assumptions.

Bootstrap

- Start from a reference tree.
- Alignment sites are sampled with replacement.
- Trees are built for each resampled dataset.
- The frequency of each node occurring in the bootstrapped trees is computed.
- Gives a statistical confidence value to each node.
- Bootstrap values of 70-75 is used as an indicator for good support.

- **Caveats**
  - Model testing – Modeltest-ng
  - Recombination Detection – GARD, Simplot



http://iqtree.cibiv.univie.ac.at/

iqtree -s Combined_HA_genbank_aln.fas -bb 1000 -st DNA -nt 4 -alrt 1000 -pre treefileout

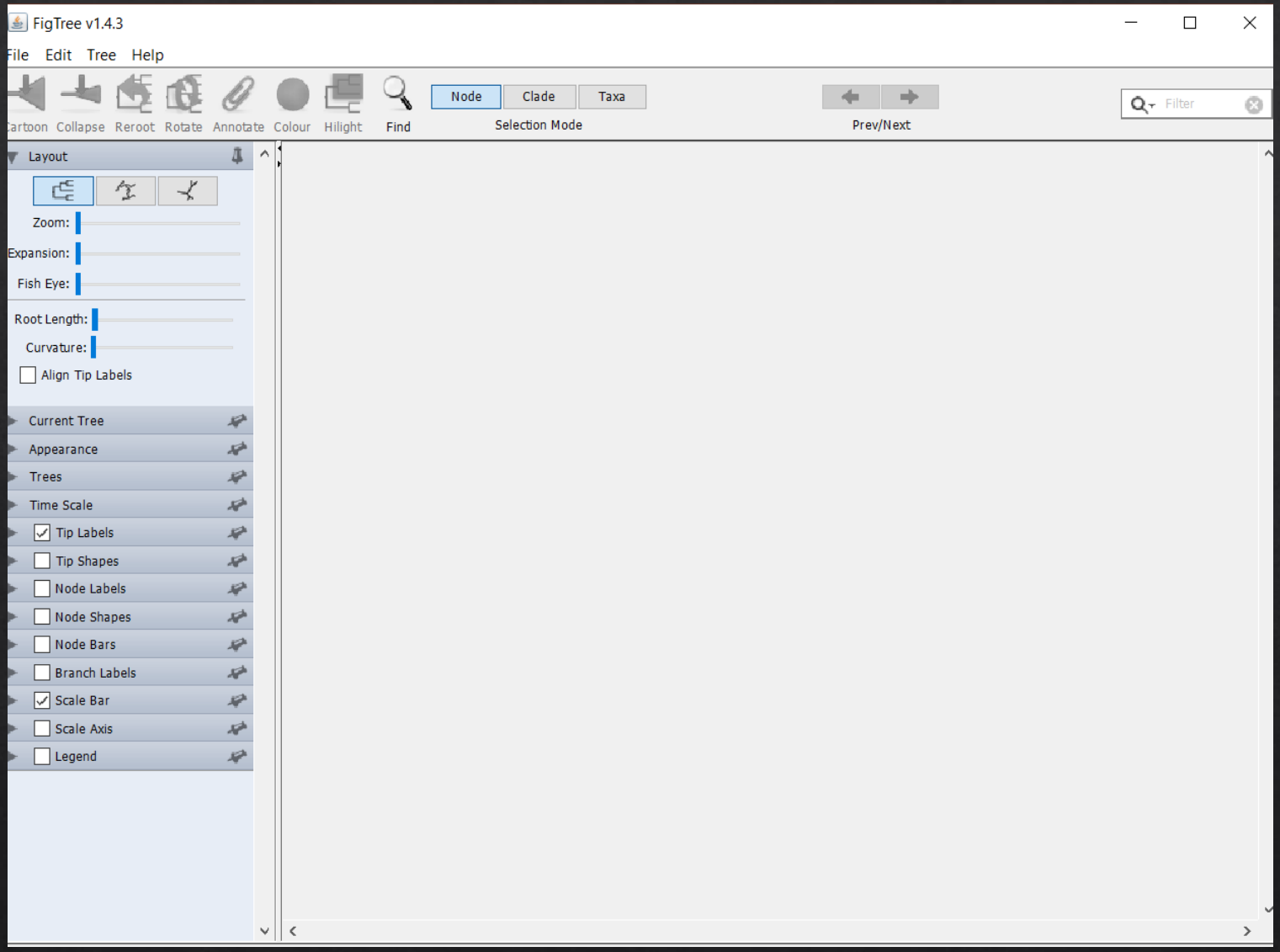
-s : Input File
-bb : ultrafast bootstrap
-st : data type
-nt : Number of threads
-alrt : SH-like approximate likelihood ratio test
-pre : Prefix for output file

Note:

- This will run with model testing included
- Output will have .treefile that has both the UF booststrap and alrt values and the .contree consensus tree
- For UF bootstrap values >95 and aLRT values >80 considered as strong support
- Traditional bootstrap can be done using –b option
- Models can be specified using –m option

Viewing and modifying a Tree file

- **Why?**
  - To visualize final phylogenetic relationships and draw inferences.
  - To create final figures for publications.


FigTree v1.4.3
File Edit Tree Help
Cartoon Collapse Reroot Rotate Annotate Colour Hilight Find
Node Clade Taxa
Selection Mode
Prev/Next
Filter
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
Align Tip Labels
Current Tree
Appearance
Trees
Time Scale
Tip Labels
Tip Shapes
Node Labels
Node Shapes
Node Bars
Branch Labels
Scale Bar
Scale Axis
Legend

figtree

Based on this data one would infer that there were multiple introductions into the hospital and only a small transmission cluster of three patients could be identified.

Questions