Sequence Alignments and Phylogeny

Genomics and Clinical Virology

Sunando Roy

sunando.roy@ucl.ac.uk

Scenario: You have determined the pathogen responsible for the outbreak above and now wish to identify a likely source of infection. Consensus sequence data from bats, rodents, humans and reference sequences are available for further analysis. Using phylogenetic analysis, what is the likely source of the infection?

Outline

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

Software

1. **Mafft** - Alignment tool
2. **MEGA/SeaVIEW**- Alignment Viewer and Editor
3. **Modeltest-ng** - Model Testing
4. **IQ-TREE** - Tree Building tool
5. **MEGA/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/)

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*. 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

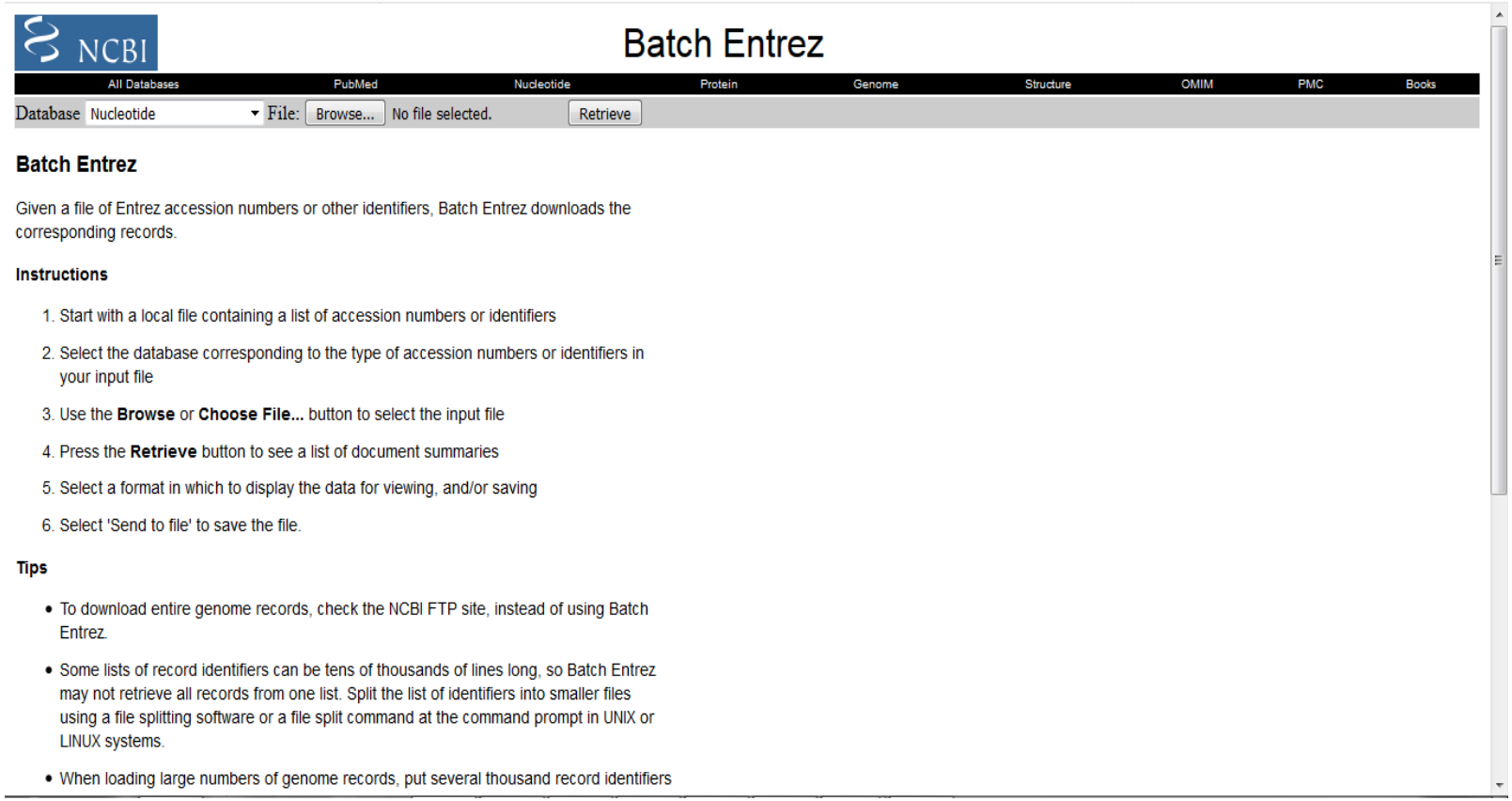
- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using NCBI e-utilities.
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.

GenBank Data Usage

GenBank Resources

GenBank Home
Submission Types
Submission Tools
Search GenBank
Update GenBank Records

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



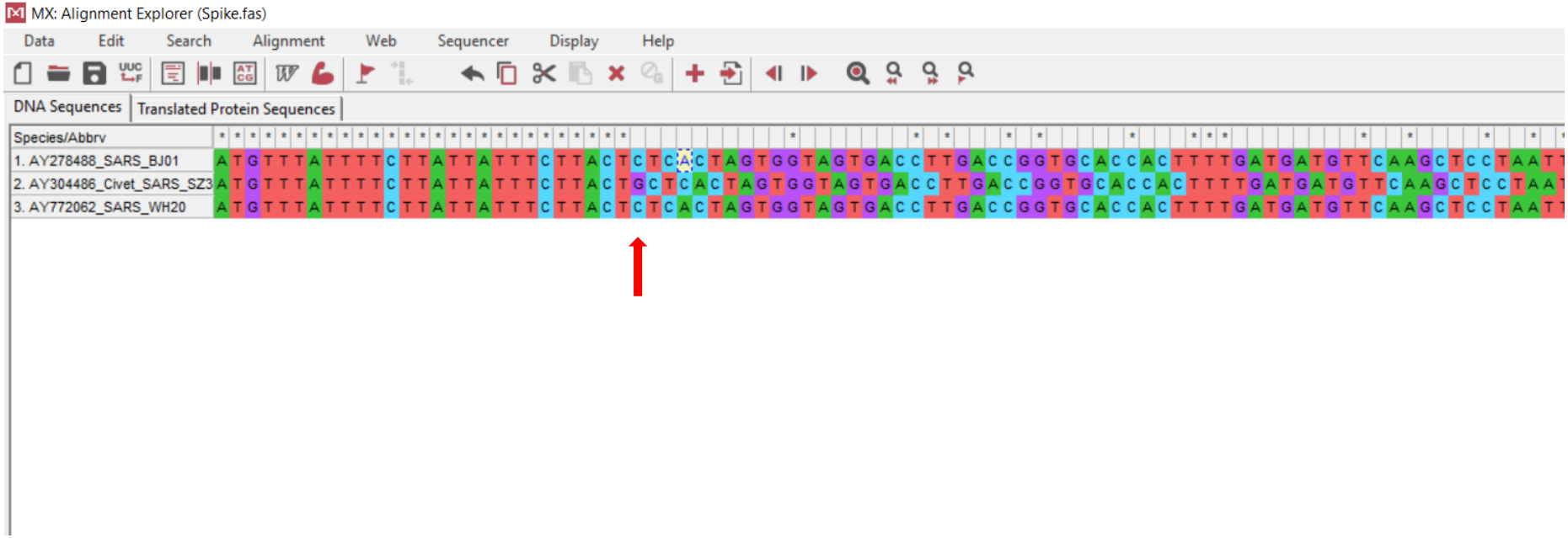
- **Alternatives**
  - Entrez E-utilities (ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/)
  - Command – Browser example

    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

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	ATGTTTATTTTCTTATTATTTCTTACT-CTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCA
2. AY304486_Civet_SARS_SZ3	ATGTTTATTTTCTTATTATTTCTTACTGCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCA
3. AY772062_SARS_WH20	ATGTTTATTTTCTTATTATTTCTTACT-CTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCA

MX: Alignment Explorer (Spike.fas)

Data Edit Search Alignment Web Sequencer Display Help

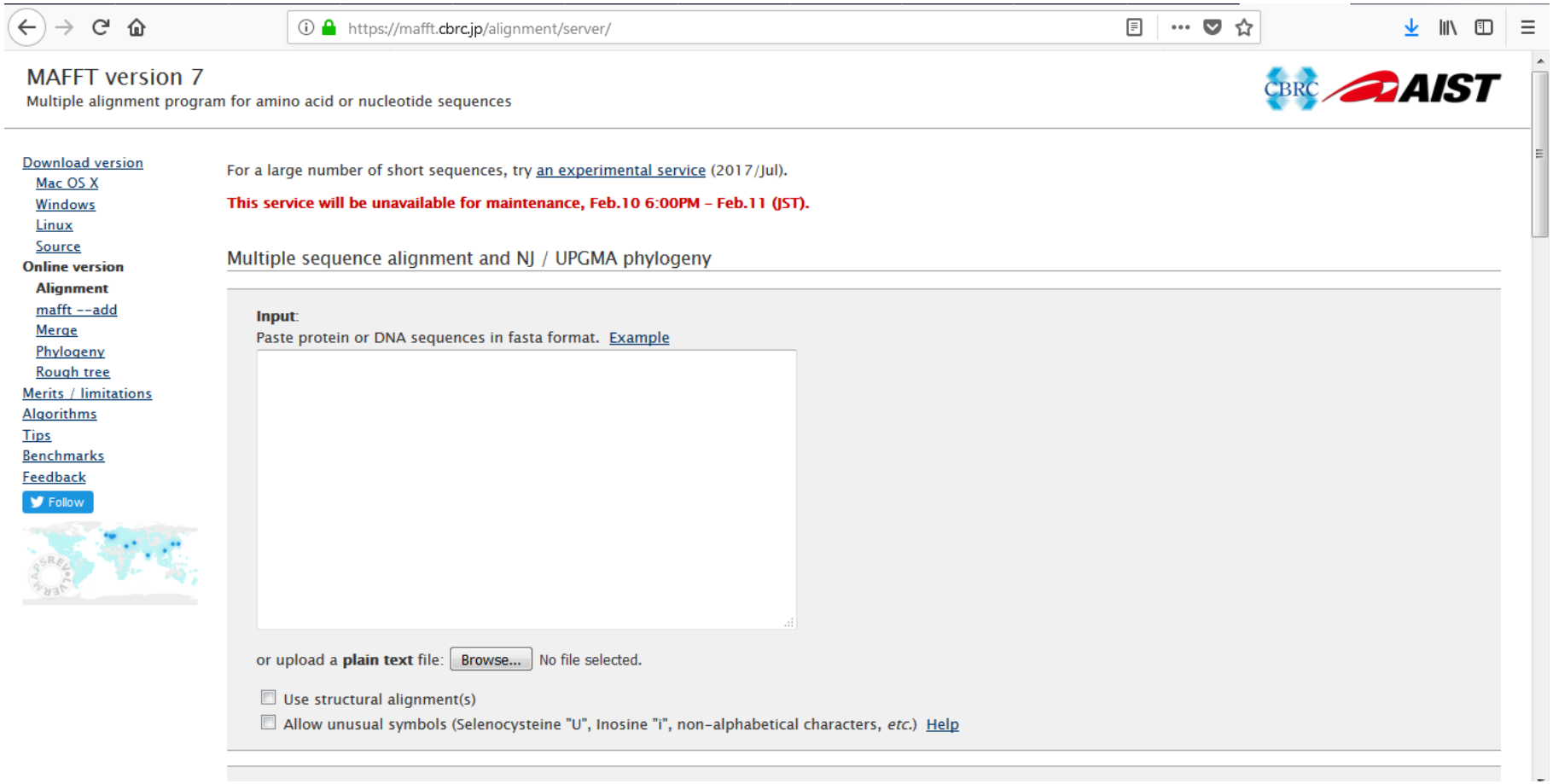
DNA Sequences | Translated Protein Sequences

Species/Abbrv	* * *
1. AY278488_SARS_BJ01	ATGTTTATTTTCTTATTATTTCTTACTCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTT
2. AY304486_Civet_SARS_SZ3	ATGTTTATTTTCTTATTATTTCTTACTCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTT
3. AY772062_SARS_WH20	ATGTTTATTTTCTTATTATTTCTTACTCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTT
4. DQ071615_Bat_Rp3	ATGAAAATTTTAATTCTTGCTTTCCTAGCTAGTCTAGCTAAAGCACAAGAAGGATGTGGCATTATCAGTCGAAAACCTCAGCCAAAAATGGCACAAGTCTCTT
5. DQ412042_Bat_Rf1	ATGAAAATTTTAATTTTTGCTTTCCTAGTTACCCTAGTTAAAGCACAAGAAGGTTGTGGCGTGATCAACCTCAGGACACAACCTAAATTAACACAAGTCTCCT
6. DQ412043_Bat_Rm1	ATGAAAGTTTTAATTTTTGCTCTGCTCTTCAGTCTAGCTAAAGCACAGGAAGGATGTGGCATTATTAGTAGAAAGCCACAGCCAAAGATGGAAAAAGTCTCAT
7. DQ648856_BtCoV_273_2005	ATGAAAATTTTAATTTTTGCTTTCCTAGTTACCCTAGTTAAAGCACAAGAAGGTTGTGGCGTGATCAACCTCAGGACACAACCTAAATTAACACAAGTCTCCT
8. DQ648857_Bat_BtCoV_279_2005	ATGAAAGTTTTAATTTTTGCTCTGCTCTTCAGTCTAGCTAAAGCACAGGAAGGATGTGGCATTATTAGTAGAAAGCCACAGCCAAAGATGGAAAAAGTCTCAT
9. FJ588686_Bat_Rs672_2006	ATGAAAGTTTTGATTGTTCTTTTATGCCTCGGCCTTGTTACTGCTCAAGATGGCTGTGGGCATATTAGCACTAAACCTCAACCATTAATGGATAAATTTTCTT
10. FJ647226_Murine_MHV-JHM	ATGCTGTTCGTCTTTATTTTACTATTACCCTCTTGTTTAGGGTATATTGGTGATTTTAGATGTATCCAGACCGTGAATTATAACGGCAATAATGCTTCTGCGC
11. JX993987_Bat_Rp_Shaanxi2011	ATGATTTTACTTCTTCTTTTCCTTTCTTCCGCTAAAGCTCAGGAAGGTTGTGGTGTTATCTCCAATAAGCCACAACGCACATTTGACCAGTACTCCTCCACTT
12. JX993988_Bat_Cp_Yunnan2011	ATGAAAATTTTCTTGTTTAGTCTGCTTTTTAGTGCTGCTTTAGCCCAGGAAGGTTGTGGCTTATTGAGTTTCAAACCACAACCTAAGCTGGCACAATTTTCATG
13. KC881005_Bat_RsSHC014	ATGAAATTGTTAGTTTTAGTTTTTGCTACTCTAGTCTCCTCTTACACTATAGAGAAGTGCCTTGATTTTGATGACCGCACTCCACCTGCAAATACTCAATTTTT
14. KC881006_Bat_Rs3367	ATGAAATTGTTAGTTTTAGTTTTTGCTACTCTAGTCTCCTCTTACACTATAGAGAAGTGCCTTGATTTTGATGACCGCACTCCACCTGCAAATACTCAATTTTT
15. KF367457_Bat_WIV1	ATGAAATTGTTAGTTTTAGTTTTTGCTACTCTAGTCTCCTCTTACACTATAGAGAAGTGCCTTGATTTTGATGACCGCACTCCACCTGCAAATACTCAATTTTT
16. KF569996_Rhinolophus_affinis_LYRa11	ATGTTTTTAACTTGTTTCATTTTATCGTTTAGTCTTTTCTGTGTGTCTGGTGATTCTATTGACACCTGTGAAACTTTTGATGATGTCTCTCCCCCACAACAAAA
17. KJ473811_BtRf-BetaCoV_JL2012	ATGTTTATTGTATTTCTCATTAGTTACACAACTTTTCTCATTCCTTACACAGCTACAACAACATGTTTCAAAGGGCCTACAACAGAGAATAAGTTGAACATATG
18. KJ473812_BtRf-BetaCoV_HeB2013	ATGAAAATTTTAATTTTTGCTTTCCTAGTTACCCTAGTTAAAGCACAAGAAGGTTGTGGCGTGATCAACCTCAAGACACAACCTATATTAACACAAGTCTCCT
19. KJ473813_BtRf-BetaCoV_SX2013	ATGAAAATTTTAATTTTTGCTTTCCTAGTTACCCTAGTTAAAGCACAAGAAGGTTGTGGCGTGATCAACCTCAAGACACAACCTACATTAACACAAGTCTCCT
20. KJ473815_BtRs-BetaCoV_GX2013	ATGAAAATTTTAATTCTTGCTTTCTTGGCTAGTTTAGCTAAAGCACAGGAAGGATGCGGCATTATCAGCAGAAAACCACAACCTAAATTGGCACAAGTATCTT
21. KP886809_Bat_YNLF_34C	ATGAAAATTTTAATTTTTGCTTTCCTAGTTACTCTAGTTAAAGCACAAGAAGGTTGTGGTGTAATTAACCTCAGGACACAACCTAAATTATCACAAGTCTCCT
22. KT444582_Bat_WIV16	ATGTTTATTTTCTTATTCTTTCTCACTCTCACTAGTGGTAGTGACCTTGAGAGTTGTACCACTTTTGATGATGTTCAAGCCCCTAATTACCCTCAACACTCTTG
23. KU182964_Bat_JTMC15	ATGTTTATTGTATTTCTCATTAGTTACACAACTTTTCTCATTCCTTATACAGCTACAACAACATGTTTCTCAGGGCCTACAATAGAGAATAAGTTGAACATATG
24. KY370043_RtRn-CoV_YN2013	ATGTTATTTATTCTTTTTATTTTACCTTCTGCTTTTGCTGTTATAGGTGATTTAAAGTGTACAACCCTCAGCATTAATGATGCCAACACTGGCGCACCTAGCA
25. KY417145_Bat_Rf4092	ATGTTTTTTCTTATTGGTTACACAGCTTTTCTCATTGGTTACACAGCTGCAACAACATGTGTCACGGGGCCTACAACAGAGAATAAGTTGAACATCTCCTCAGG
26. KY417146_Bat_Rs4231	ATGTTTATTTTCTTATTCTTTCTCACTCTCACTAGTGGTAGTGACCTTGAGAGTTGTACCACTTTTGATGATGTTCAAGCCCCTAATTACCCTCAACACTCTTG
27. KY417150_Bat_Rs4874	ATGTTTATTTTCTTATTCTTTCTCACTCTCACTAGTGGTAGTGACCTTGAGAGTTGTACCACTTTTGATGATGTTCAAGCCCCTAATTACCCTCAACACTCTTT
28. MG772933_Bat-SL-CoVZC45	ATGTTGTTTTTCTTGTTTCTTCAGTTCGCCTTAGTAAACTCCCAGTGTGTTAACTTGACAGGCAGAACCCCACTCAATCCCAATTATACTAATTCTTCACAAAG
29. MG772934_Bat-SL-CoVZXC21	ATGTTGTTTTTCTTGTTTCTTCAGTTCGCCTTAGTAAACTCCCAGTGTGATTTGACAGGTAGAACTCCACTCAATCCCAATTATACTAATTCTTCACAAAGAG
30. MK211374_BtRl-BetaCoV_SC2018	ATGATTATACTTCTTCTTTTTCTTTCCTCAGCTACTGCGCAGGAAGGTTGTGGTGTGATATCCAACAAACCACAGCGCACGTTTGACCAATACTCTTCTACTCG
31. MK211375_BtRs-BetaCoV_YN2018A	ATGAAAATCTTAATTTTTGCTTTTCTAGTTACGCTAGTTGAAGCACAGGAAGGATGCGGCATTATCAGTAGAAAACCGCAGCCAAAAATGGCACAAGTCTCCT
32. MK211376_BtRs-BetaCoV_YN2018B	ATGAAATTGTTAGTTTTAGTTTTTGCTACTTTGGTCTCTCTTACACTATAGAGAAGTGCACTGACTTTGATGATCGCACTCCACCTTCAAACACTCAATTTCT
33. MK211377_BtRs-BetaCoV_YN2018C	ATGAAAATCTTAATTTTTGCTTTTCTAGTTACGCTAGTTAAAGCACAGGAAGGATGCGGCATTATCAGCAAAAAACCGCAGCCAAAAATGGCACAAGTTTCCT
34. MK211378_BtRs-BetaCoV_YN2018D	ATGAAAGTTTTGATTGTTCTTTTATGCCTCGGCCTTGTTACTGCTCAAGATGGCTGTGGGCATATTAGCACTAAACCTCAACCATTATTGGATAAATTTTCTT
35. MK211379_BtRt-BetaCoV_GX2018	ATGTGGGCGATATGGATATTGTTAATATTGCCAACAGTACTTGCTCGCCCTAACTGTGTTACTTTGAGTCTACCTCATATTGAACACAATGTAACTTATCAGT
36. Wuhan-Hu-1	ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCAATTACCCCCTGCATACACTAATTCTTTCACACGTG

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.

Code: $ mafft --maxiterate 1000 --localpair ~/Sunando/Spike.fas > outputfile.fas



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

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 time of transmission.

Tree Building

Seq 1 – ATTGCAAT

Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

Tree Building

Seq 1 – ATTGCAAT

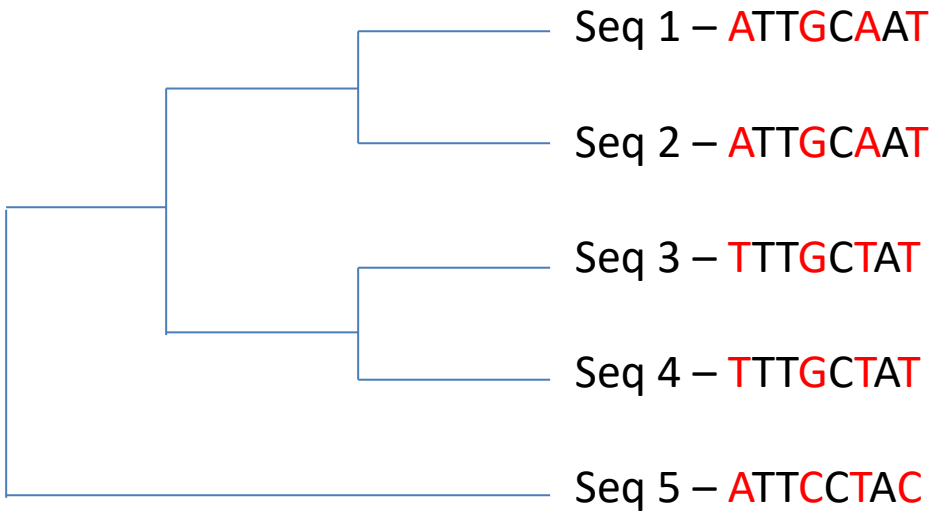
Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

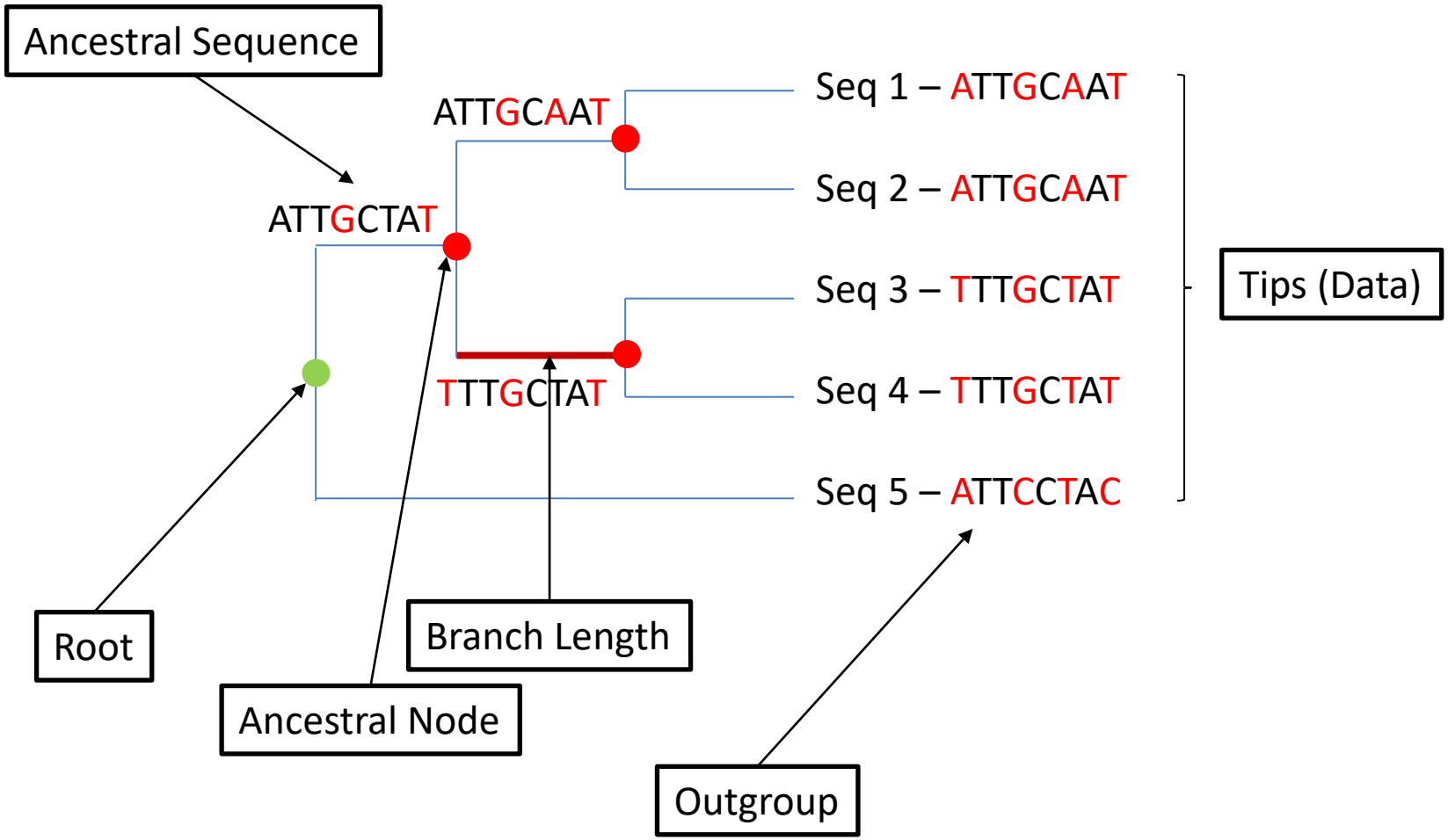
Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

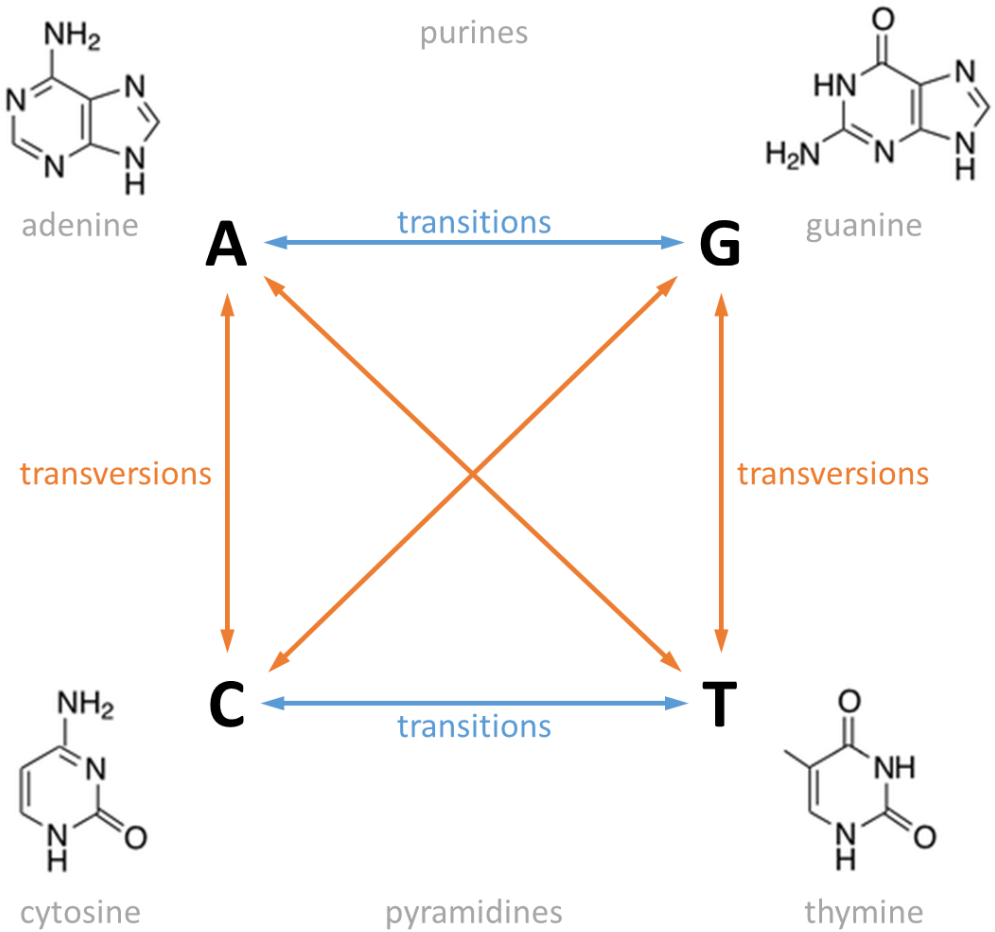
Tree Building



Tree Building



Evolutionary models - DNA



Jukes Cantor

Complexity

General Time Reversible

Evolutionary models - Proteins

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Model	Training data	References
General models:		
JTT	—	Jones *et al.* (1992)
LG	—	Le and Gascuel (2008)
PAM (Dayhoff)	—	Dayhoff *et al.* (1978)
PMB	—	Veerassamy *et al.* (2003)
VT	—	Müller and Vingron (2000)
WAG	—	Whelan and Goldman (2001)
Specialized models:		
HIVb	HIV (eight proteins)	Nickle *et al.* (2007)
rtREV	retroelement *pol*	Dimmic *et al.* (2002)

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

- Tree Building Algorithms
  - Neighbor Joining – BioNJ
  - Maximum Parsimony – MEGA
  - Maximum Likelihood – PhyML, RAxML,IQ-TREE
  - 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 Parsimony

- Character based tree estimation
- Evaluates multiple tree topologies.
- Scores the best tree on minimum number of character changes required to explain the data.
- Does not use evolutionary models of substitution.
- Does not perform well over longer evolutionary time scales.
- Suffers from effects of Long Branch Attraction.

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

Code:

```
$ iqtree -s ~/Sunando/Spike_aln.fas -bb 1000 -st DNA -nt 4 -alrt 1000 -pre treeoutfile
```

-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

iqtree.cibiv.univie.ac.at

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 33%

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

Tree Inference | Model Selection | Analysis Results

For a quick start, take a look at the tutorial for the IQ-TREE web server.
Please visit the IQ-TREE homepage for more information or if you want to download the main software.
Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file : Browse... Show example >

Use example alignment: Yes ?

Sequence type: Auto-detect DNA Protein Codon DNA->AA Binary Morphology ?

Partition file: This field is optional. Browse... Show example >

Partition type: Edge-linked Edge-unlinked ?

Substitution Model Options

Substitution model: Auto ?

FreeRate heterogeneity: Yes [+R]

Rate heterogeneity: Gamma [+G] Invar. sites [+I] ?

#rate categories: 4

State frequency: Empirical (from data) AA model (from matrix) ML-optimized Codon F1x4 Codon F3x4

Ascertainment bias correction: Yes [+ASC] ?

Branch Support Analysis

Bootstrap analysis: None Ultrafast Standard ?

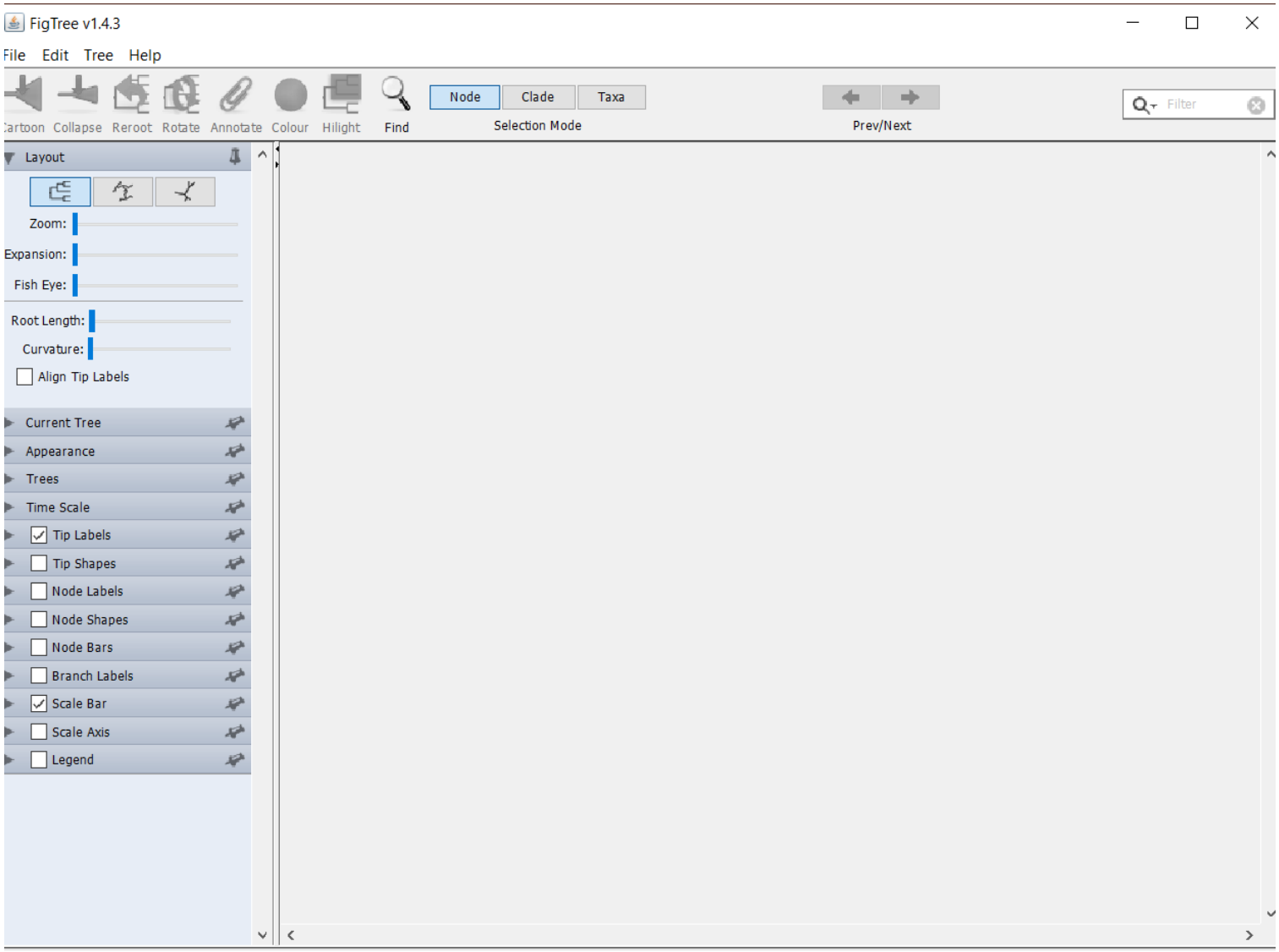
Number of bootstrap alignments: 1000

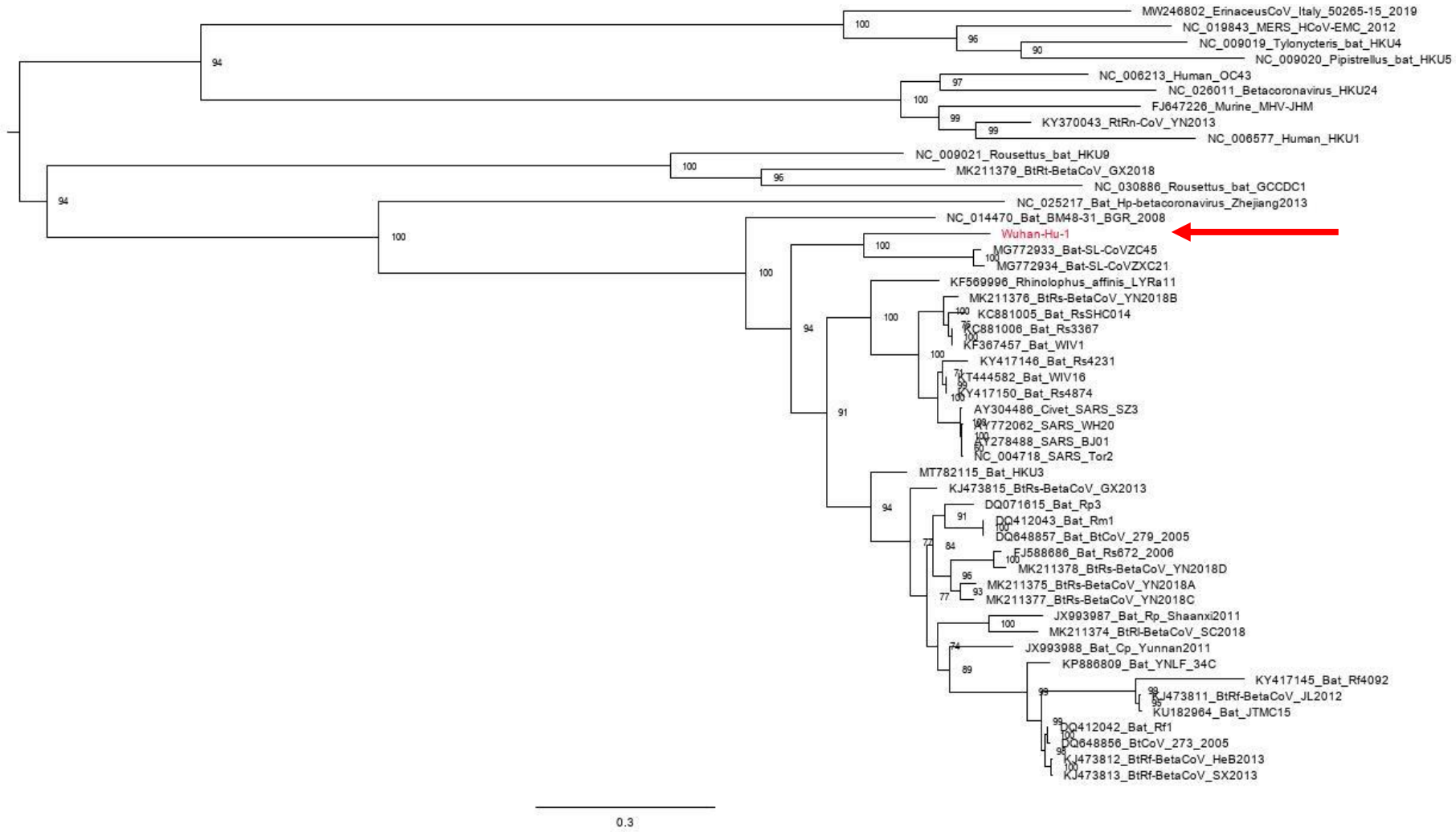
http://iqtree.cibiv.univie.ac.at/

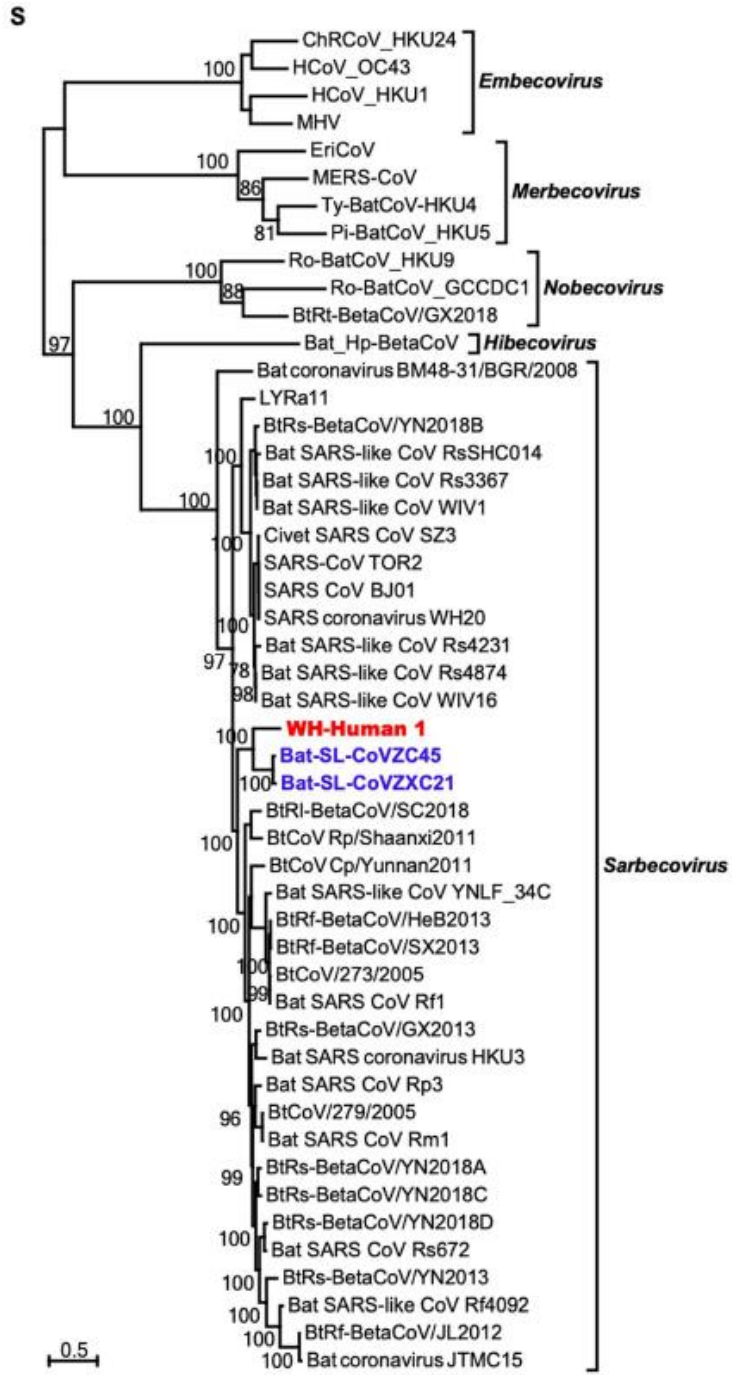
Viewing and modifying a Tree file

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

Code: $ figtree




MW246802_ErinaceusCoV_Italy_50265-15_2019
NC_019843_MERS_HCoV-EMC_2012
NC_009019_Tylonycteris_bat_HKU4
NC_009020_Pipistrellus_bat_HKU5
NC_006213_Human_OC43
NC_026011_Betacoronavirus_HKU24
FJ647226_Murine_MHV-JHM
KY370043_RtRn-CoV_YN2013
NC_006577_Human_HKU1
NC_009021_Rousettus_bat_HKU9
MK211379_BtRt-BetaCoV_GX2018
NC_030886_Rousettus_bat_GCCDC1
NC_025217_Bat_Hp-betacoronavirus_Zhejiang2013
NC_014470_Bat_BM48-31_BGR_2008
Wuhan-Hu-1
MG772933_Bat-SL-CoVZC45
MG772934_Bat-SL-CoVZXC21
KF569996_Rhinolophus_affinis_LYRa11
MK211376_BtRs-BetaCoV_YN2018B
KC881005_Bat_RsSHC014
KC881006_Bat_Rs3367
KF367457_Bat_WIV1
KY417146_Bat_Rs4231
KT444582_Bat_WIV16
KY417150_Bat_Rs4874
AY304486_Civet_SARS_SZ3
AY772062_SARS_WH20
AY278488_SARS_BJ01
NC_004718_SARS_Tor2
MT782115_Bat_HKU3
KJ473815_BtRs-BetaCoV_GX2013
DQ071615_Bat_Rp3
DQ412043_Bat_Rm1
DQ648857_Bat_BtCoV_279_2005
FJ588686_Bat_Rs672_2006
MK211378_BtRs-BetaCoV_YN2018D
MK211375_BtRs-BetaCoV_YN2018A
MK211377_BtRs-BetaCoV_YN2018C
JX993987_Bat_Rp_Shaanxi2011
MK211374_BtRl-BetaCoV_SC2018
JX993988_Bat_Cp_Yunnan2011
KP886809_Bat_YNLF_34C
KY417145_Bat_Rf4092
KJ473811_BtRf-BetaCoV_JL2012
KU182964_Bat_JTMC15
DQ412042_Bat_Rf1
DQ648856_BtCoV_273_2005
KJ473812_BtRf-BetaCoV_HeB2013
KJ473813_BtRf-BetaCoV_SX2013
0.3



Based on this data one would infer that the isolated virus most likely lies within the Sarbecovirus group which includes the original SARS coronavirus but is quite distinct from it. The closest related species both are from bats which could suggest a potential origin

Nature, *579*(7798), 265–269.

Questions