Tutorial:

Multi-locus molecular phylogenetics

ballesterus

May 23, 2021

Contents

1 Goals

The goal of this tutorial is to provide hands-on experience with the procedures and decisions involved in phylogentic inference from multiple genetic loci.

2 Required software

There is a large list of tools/software for all the steps of phylogenetic inference. The computational challenges of phylogenetic inference remain a dynamic area of research and new methods and tools are constantly beig developed. In the end the specifics of the project, practice, experience would best inform the selection of the right tool(s) to use.

For this exercise we will be using:

- MEGA: This is an integrative suite for molecular evolutionary biology. MEGA is feature rich and it is a project in active development [6] and academic licenses freely available for all OS at `https://www.megasoftware.net`

- Aliview: This is a free molecular sequence viewer and editor. It requires java and it run in Linux/Windows/MacOS operating system. It includes MUSCLE as a default multiple sequence aligner and can be further customized to use other programs. `https://ormbunkar.se/aliview/`

- MACSE: This java program is designed to align nucleotide coding sequences as both aminoacid and nucleotides. The program is also available online at the following link `https://mbb.univ-montp2.fr/MBB/subsection/softExec.php?soft=macse2` or can be dowloaded locally at

  - trimAL This tool uses an explicit algorithms to remove gappy or unreliable regions from a multiple sequence alignment. [1] can be downlowed from

- IQTREE This program estimates phylogenetic trees from a multiple sequence alignment [4]. This program has many useful and convenient features, it also incorporate model selection and produces detailed, easy to read outputs.

- Figtree: It is also based in Java, portable across OS and convenient for viewing, annotating and printing trees in NEWICK and NEXUS formats.

3 Exercise: Multilocus phylogeny of bears

The dataset in this excercise is loosely based on [3]. The purpose of the excercise is not to test the published result but to demonstrate and practice the routines and steps for inferring phylogenies from multiple loci.

File	Locus	Notes
cytb.fasta	Cythochrome oxidase B	Mitochondrial protein coding sequence (CDS)
fes.fasta	Feline sarcoma proto-oncogene	Nuclear protein coding sequence
irbp.fasta	interphotoreceptor retinoid binding protein (irbp)	Nuclear protein coding sequence
r12S	ribosomal subunit 12S	Mitochondrial ribosomal gene

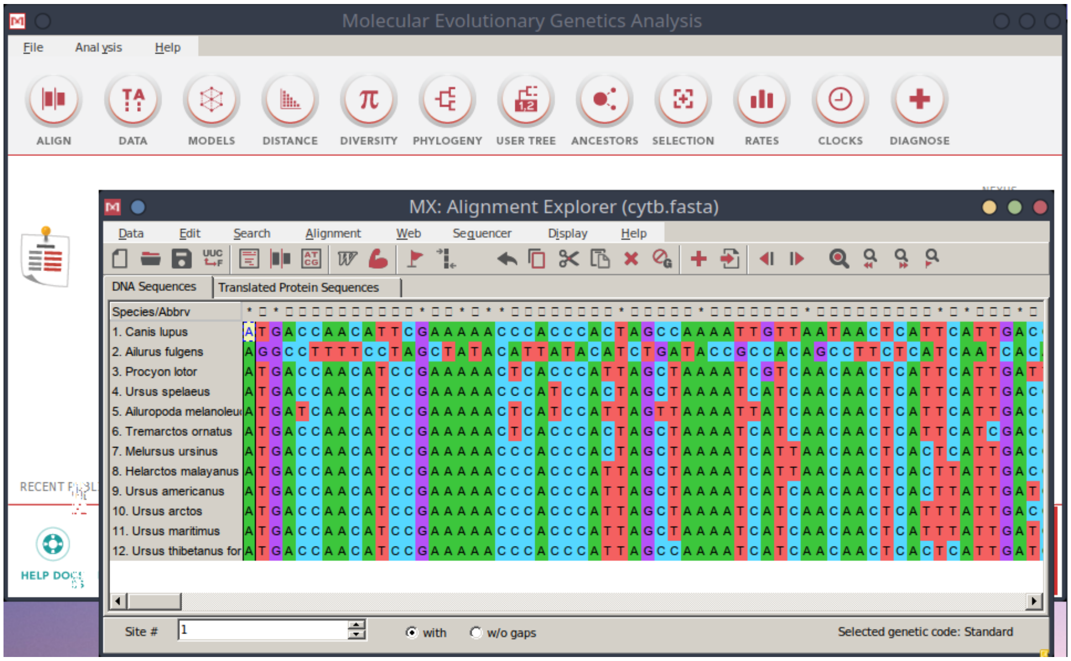
3.1 Download excercise data files

It is good practice to organize data files of a project in a **known** directory in your harddrive. Many programs will "manage" data for you but as a consequence you may not know where your files are!

3.2 The one package approach using MEGA

1. Launch MEGA

2. Load one of cytb.fasta file into the program. Use the **DATA** icon > **Open-File/Session** and navigate to the corresponding file.



Species/Abbrv	
1. Canis lupus	ATGACCAACATTCGAAAAACCCACCCACTAGCCAAAATTGTTAATAACTCATTCATTGAC
2. Ailurus fulgens	AGGCCTTTTCCTAGCTATACATTATACATCTGATACCGCCACAGCCTTCTCATCAATCAC
3. Procyon lotor	ATGACCAACATCCGAAAAACTCACCCATTAGCTAAAATCGTCAACAACTCATTCATTGAT
4. Ursus spelaeus	ATGACCAACATCCGAAAAACCCATCCACTAGCTAAAATCATCAACAACTCATTCATTGAC
5. Ailuropoda melanoleu	ATGATCAACATCCGAAAAACTCATCCATTAGTTAAAATTATCAACAACTCATTCATTGAC
6. Tremarctos ornatus	ATGACCAACATCCGAAAAACTCACCCACTAGCTAAAATCATCAACAACTCATTCATCGAC
7. Melursus ursinus	ATGACCAACATCCGAAAAACCCACCCACTAGCTAAAATCATTAACAACTCACTCATTGAC
8. Helarctos malayanus	ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATTAACAACTCACTTATTGAC
9. Ursus americanus	ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTTATTGAT
10. Ursus arctos	ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGAC
11. Ursus maritimus	ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGAT
12. Ursus thibetanus for	ATGACCAACATCCGAAAAACCCACCCATTAGCCAAAATCATCAACAACTCACTCATTGAT

Site # 1 ● with ○ w/o gaps Selected genetic code: Standard

3. Select align using MUSCLE algorithm. Note that this option opens the menu for changing specific alignment parameters for MUSCLE [2]. These parameters can be tuned for specific problems but default parameters usually work fine, refer to the manual to .

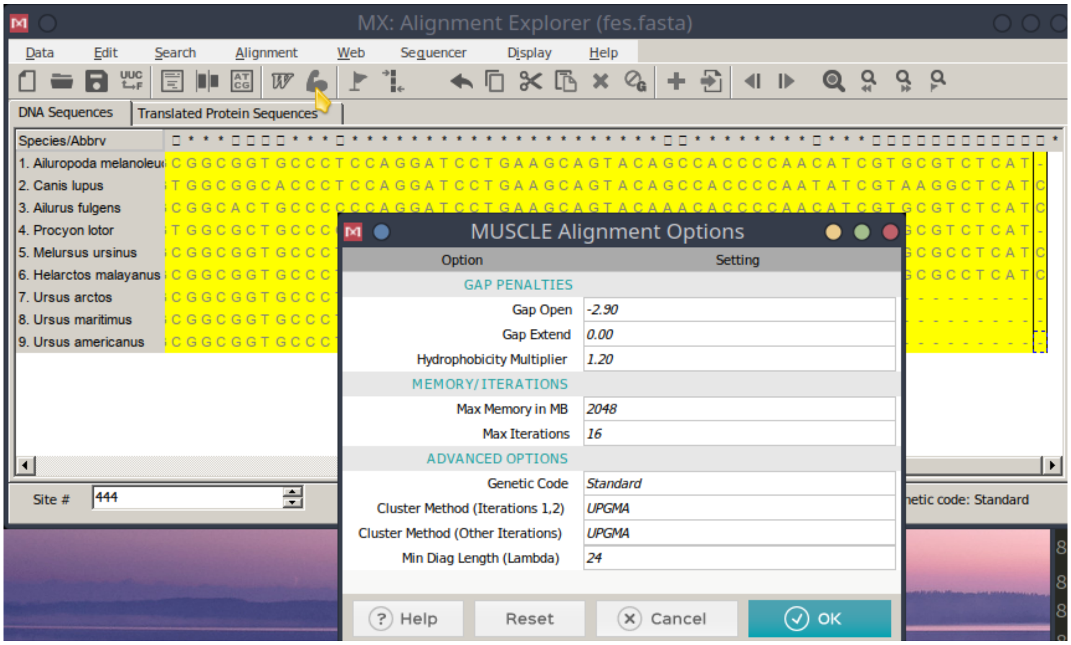
4. Visually inspect the alignment. Regions of large incongruence with many gaps may be indicative of erroneous alignment; specially if they involve few of the sequences. Tips to trouble shoot alignment issues:

   - Verify gene homology via BLAST search
   - Verify the fragments are overlapping (global homology)
   - Verify strand direction

5. At opening you will be ask to either "Align" or "Analyze" the file. Select Align. This will open MEGA'S Alignment explorer.

6. Decide if you want to manually trim certain regions. This procedure must be justified and is useful to explore untrimmed analyses first.

7. Save your finished alignment in the working directory. **FILE > EXPORT ALIGMENT > FASTA FORMAT** . Use a meaningful name and avoid spaces and special characters. Use standard sequence formats (FASTA, PHYLIP, NEXUS) over program specific data formats.



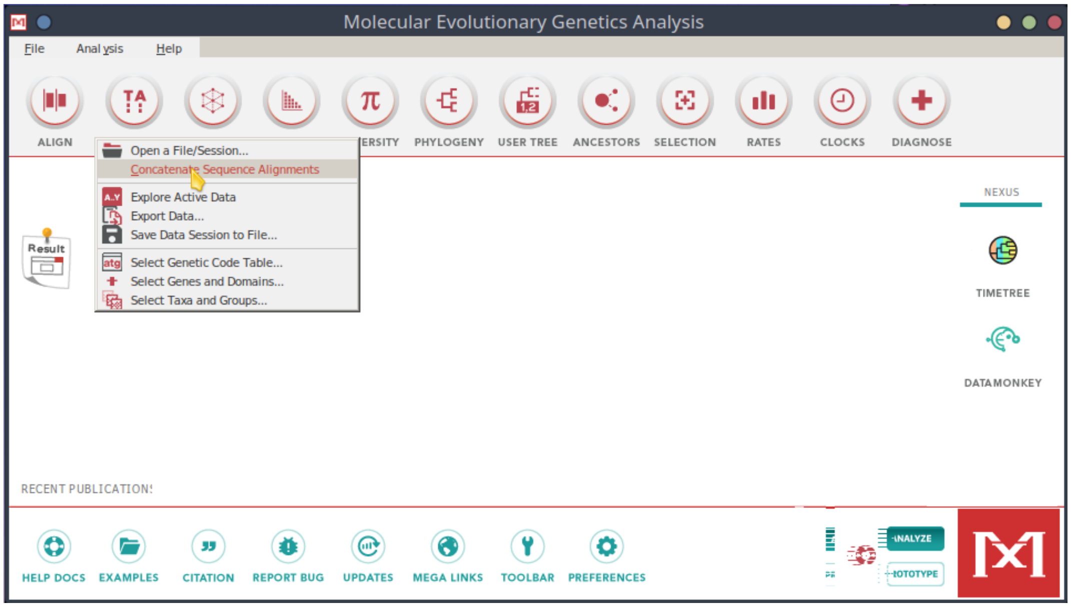
Option	Setting
GAP PENALTIES	
Gap Open	-2.90
Gap Extend	0.00
Hydrophobicity Multiplier	1.20
MEMORY/ITERATIONS	
Max Memory in MB	2048
Max Iterations	16
ADVANCED OPTIONS	
Genetic Code	Standard
Cluster Method (Iterations 1,2)	UPGMA
Cluster Method (Other Iterations)	UPGMA
Min Diag Length (Lambda)	24

8. Repeat steps 2 to 8 for each of the sequences in the data folder.

9. Copy all the aligned "finished" sequences in a new folder.

10. Use the **DATA icon** and select the option concatenate. Note: For correct concatenation the species/OTUS names must be **exactly** the same across files(spaces, underscores and punctuation marks count as differences).

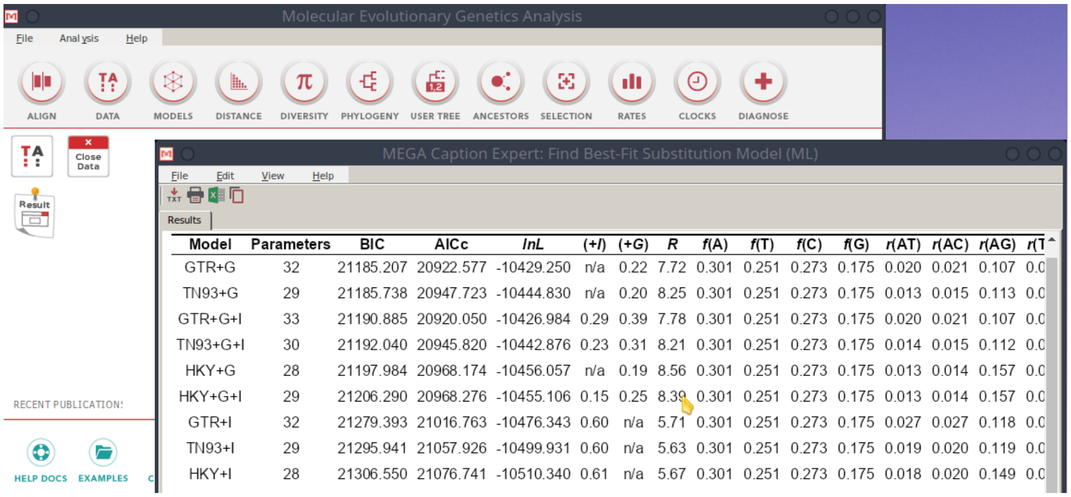
11. Navigate to the data folder and select the folder containing the finished alignments.



12. Select the appropiate data type (Nucleotide/Aminoacid) and hit **OK**

13. From the **MODELS button** > **find best DNA/Protein model (ML)...**. Leave default options unmodified in the Models selection menu and hit OK.

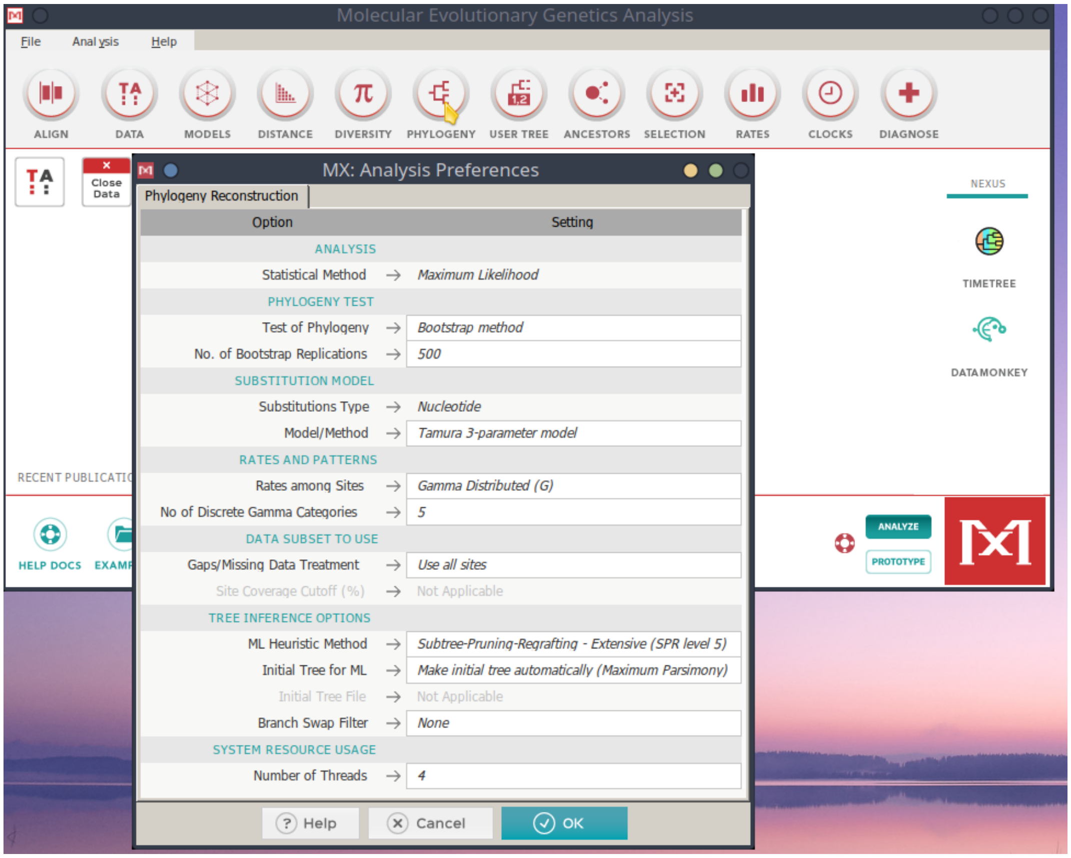
14. Models with the lowest BIC scores (Bayesian Information Criterion) shown at the top of the list are considered ones that best fit the data. Take note of the best model identified in this analysis.



Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(T
GTR+G	32	21185.207	20922.577	-10429.250	n/a	0.22	7.72	0.301	0.251	0.273	0.175	0.020	0.021	0.107	0.0
TN93+G	29	21185.738	20947.723	-10444.830	n/a	0.20	8.25	0.301	0.251	0.273	0.175	0.013	0.015	0.113	0.0
GTR+G+I	33	21190.885	20920.050	-10426.984	0.29	0.39	7.78	0.301	0.251	0.273	0.175	0.020	0.021	0.107	0.0
TN93+G+I	30	21192.040	20945.820	-10442.876	0.23	0.31	8.21	0.301	0.251	0.273	0.175	0.014	0.015	0.112	0.0
HKY+G	28	21197.984	20968.174	-10456.057	n/a	0.19	8.56	0.301	0.251	0.273	0.175	0.013	0.014	0.157	0.0
HKY+G+I	29	21206.290	20968.276	-10455.106	0.15	0.25	8.39	0.301	0.251	0.273	0.175	0.013	0.014	0.157	0.0
GTR+I	32	21279.393	21016.763	-10476.343	0.60	n/a	5.71	0.301	0.251	0.273	0.175	0.027	0.027	0.118	0.0
TN93+I	29	21295.941	21057.926	-10499.931	0.60	n/a	5.63	0.301	0.251	0.273	0.175	0.019	0.020	0.119	0.0
HKY+I	28	21306.550	21076.741	-10510.340	0.61	n/a	5.67	0.301	0.251	0.273	0.175	0.018	0.020	0.149	0.0

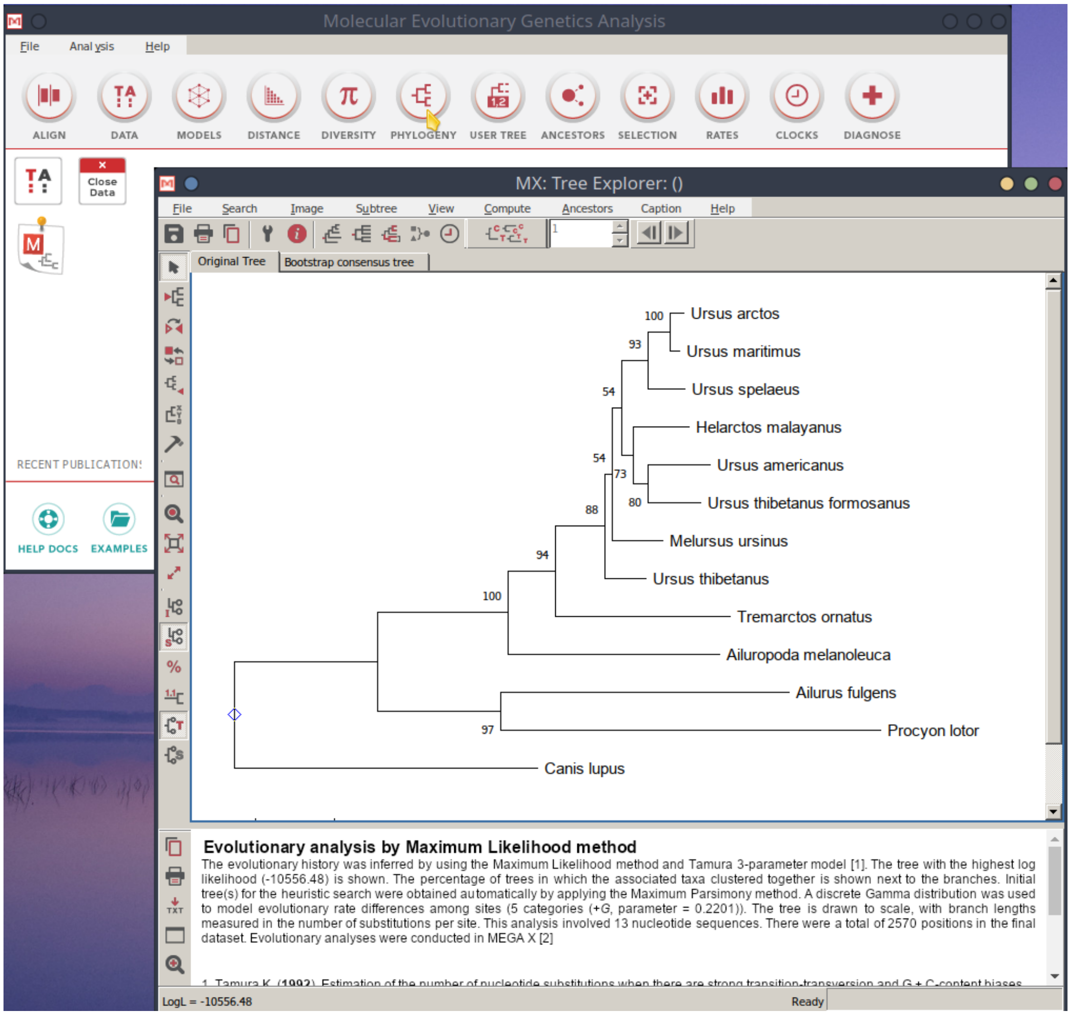
15. Use the **PHYLOGENY menu** and select Construct/test Maximum Likelihood tree analysis. Edit the Substitution MODEL and RATES AND PATTERNS menus to match the MODEL SELECTION results from step 13. By default it runs 500 bootstraps. for this exercise change that number to 100, but 500 or more are recommended for production analyses.

16. From the TREE INFERENCE OPTION make sure SPR level 5 is selected and them hit **OK**



Option		Setting
ANALYSIS		
Statistical Method	→	Maximum Likelihood
PHYLOGENY TEST		
Test of Phylogeny	→	Bootstrap method
No. of Bootstrap Replications	→	500
SUBSTITUTION MODEL		
Substitutions Type	→	Nucleotide
Model/Method	→	Tamura 3-parameter model
RATES AND PATTERNS		
Rates among Sites	→	Gamma Distributed (G)
No of Discrete Gamma Categories	→	5
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	Use all sites
Site Coverage Cutoff (%)	→	Not Applicable
TREE INFERENCE OPTIONS		
ML Heuristic Method	→	Subtree-Pruning-Regrafting - Extensive (SPR level 5)
Initial Tree for ML	→	Make initial tree automatically (Maximum Parsimony)
Initial Tree File	→	Not Applicable
Branch Swap Filter	→	None
SYSTEM RESOURCE USAGE		
Number of Threads	→	4

17. When the analyses is finished it will open a tree explorer menu.



Evolutionary analysis by Maximum Likelihood method

The evolutionary history was inferred by using the Maximum Likelihood method and Tamura 3-parameter model [1]. The tree with the highest log likelihood (-10556.48) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying the Maximum Parsimony method. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 0.2201)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 13 nucleotide sequences. There were a total of 2570 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

1. Tamura K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases.

LogL = -10556.48 Ready

18. From Tree Explorer menu select File > Export current tree (Newick). And from the export options menu, check "Branch length" and "Bootstrap options"

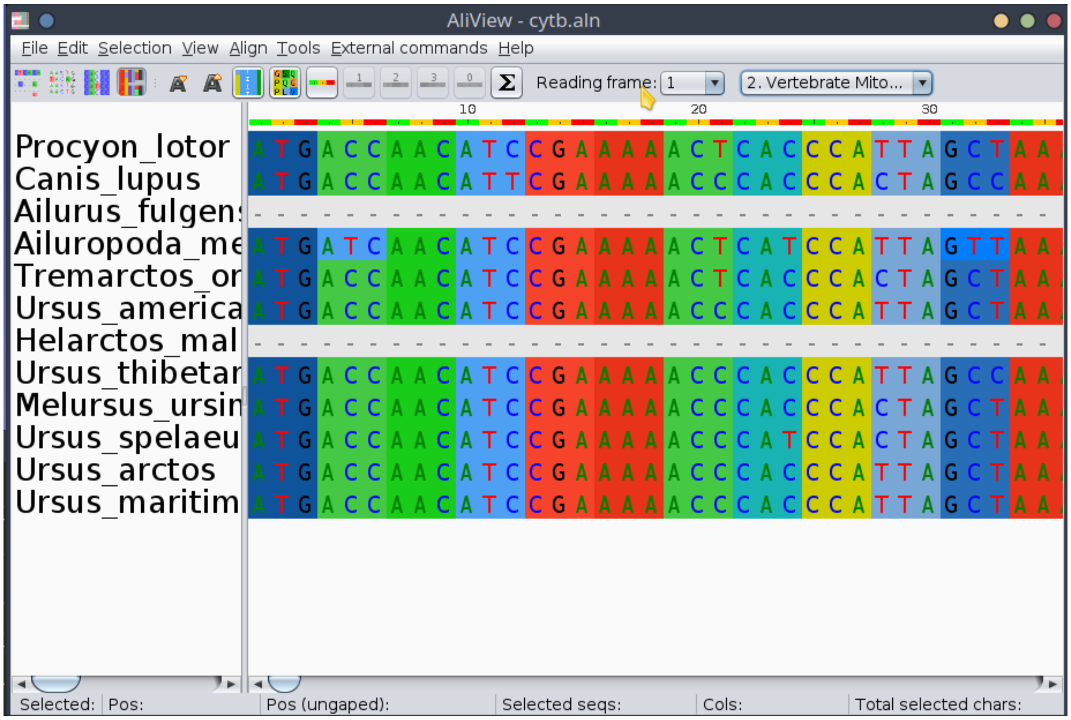
19. The parenthetical NEWICK tree can be opened by other programs such as Figtree.

3.3 Multiple programs approach

3.3.1 Multiple sequence alignment: Aliview align and edit FASTA files

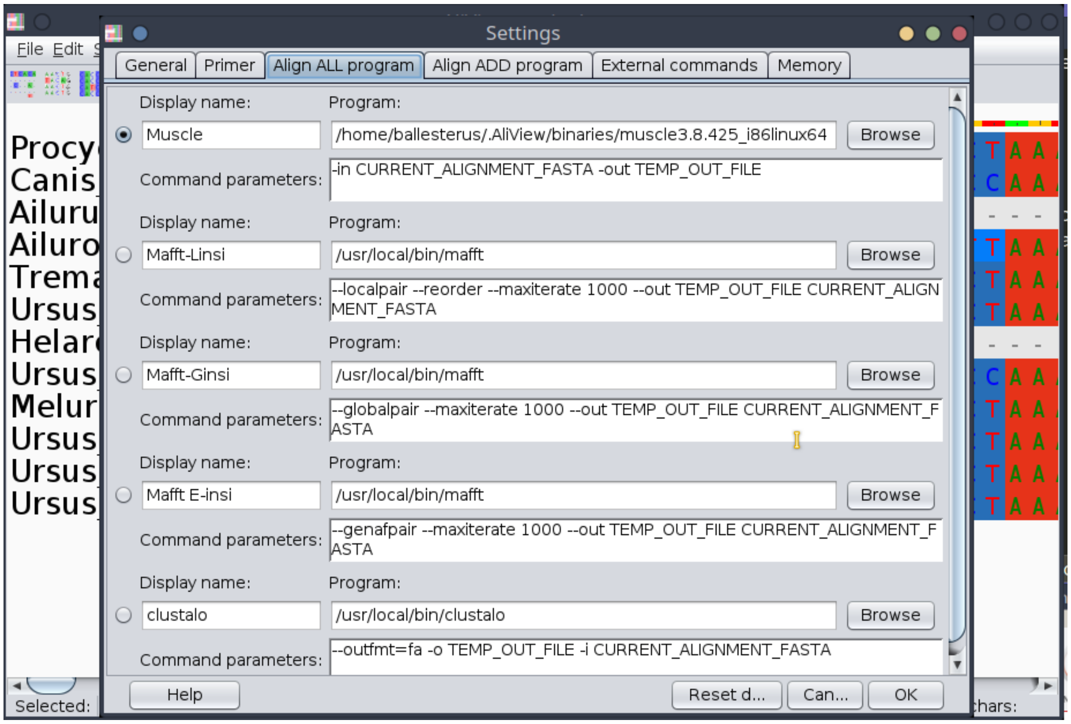
1. Fasta files can be loaded in Aliview by the file open menu or by dragging the file into the program window or by pasting the sequences in the window.

2. Inspection of reading frame is easy using the icons



```
Procyon_lotor   ATGACCAACATCCGAAAAACTCACCCATTAGCTAA
Canis_lupus     ATGACCAACATTCGAAAAACCCACCCACTAGCCAA
Ailurus_fulgen  -----------------------------------
Ailuropoda_me   ATGATCAACATCCGAAAAACTCATCCATTAGTTAA
Tremarctos_or   ATGACCAACATCCGAAAAACTCACCCACTAGCTAA
Ursus_america   ATGACCAACATCCGAAAAACCCACCCATTAGCTAA
Helarctos_mal   -----------------------------------
Ursus_thibetar  ATGACCAACATCCGAAAAACCCACCCATTAGCCAA
Melursus_ursin  ATGACCAACATCCGAAAAACCCACCCACTAGCTAA
Ursus_spelaeu   ATGACCAACATCCGAAAAACCCATCCACTAGCTAA
Ursus_arctos    ATGACCAACATCCGAAAAACCCACCCATTAGCTAA
Ursus_maritim   ATGACCAACATCCGAAAAACCCACCCATTAGCTAA
```

3. Muscle is provided by default but additional programs and tools can be configured.



● Display name: Muscle
Program: /home/ballesterus/.AliView/binaries/muscle3.8.425_i86linux64
Command parameters: -in CURRENT_ALIGNMENT_FASTA -out TEMP_OUT_FILE

○ Display name: Mafft-Linsi
Program: /usr/local/bin/mafft
Command parameters: --localpair --reorder --maxiterate 1000 --out TEMP_OUT_FILE CURRENT_ALIGNMENT_FASTA

○ Display name: Mafft-Ginsi
Program: /usr/local/bin/mafft
Command parameters: --globalpair --maxiterate 1000 --out TEMP_OUT_FILE CURRENT_ALIGNMENT_FASTA

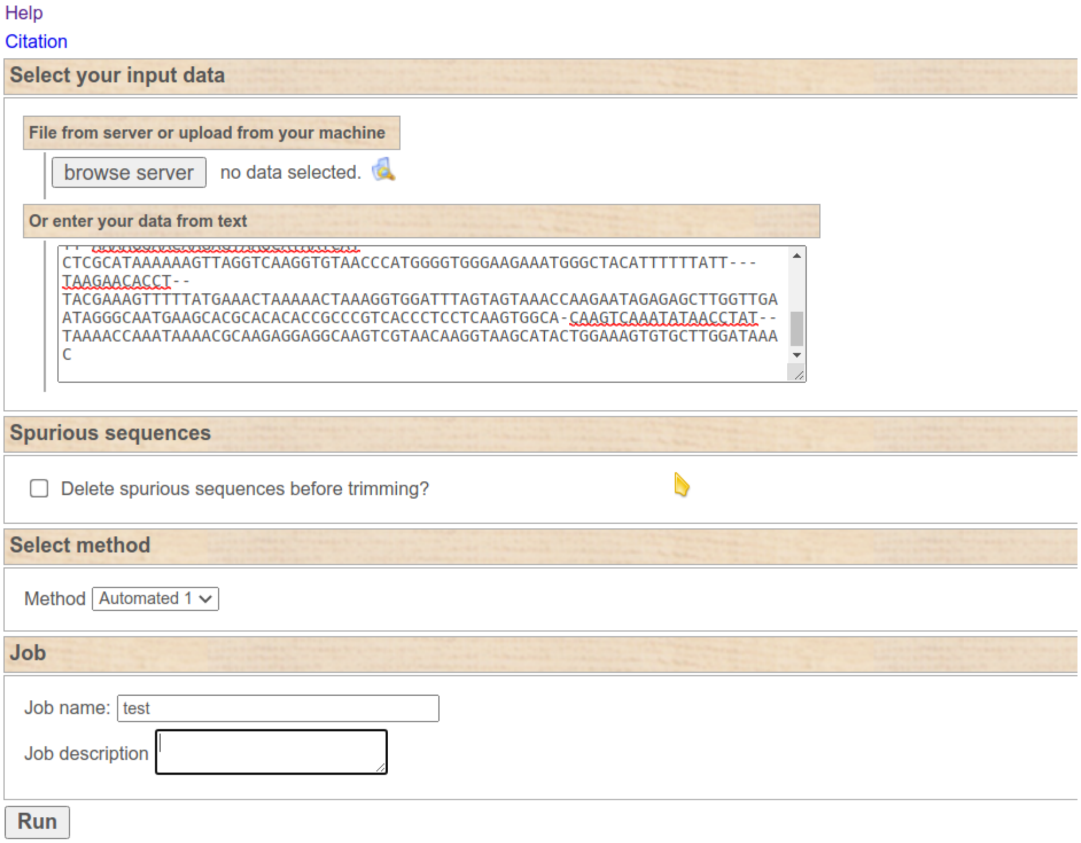
○ Display name: Mafft E-insi
Program: /usr/local/bin/mafft
Command parameters: --genafpair --maxiterate 1000 --out TEMP_OUT_FILE CURRENT_ALIGNMENT_FASTA

○ Display name: clustalo
Program: /usr/local/bin/clustalo
Command parameters: --outfmt=fa -o TEMP_OUT_FILE -i CURRENT_ALIGNMENT_FASTA

3.3.2 Trimming and sanitation

To avoid subjective editing of the data, there are dedicated tools that apply explicit criteria for removing potentially spurious alignment sites. GBLOCKS [5] and Trimal [1] are among the most popular. You can try trimal features at: `http://phylemon2.bioinfo.cipf.es/index.html`

1. Login as anonymous user
2. Select UTILITIES > ALIGNMENT UTILITIES > Trimal 1.3



Help
Citation

Select your input data

File from server or upload from your machine

browse server   no data selected.

Or enter your data from text

```
CTCGCATAAAAAAGTTAGGTCAAGGTGTAACCCATGGGGTGGGAAGAAATGGGCTACATTTTTTATT---
TAAGAACACCT--
TACGAAAGTTTTTATGAAACTAAAAACTAAAGGTGGATTTAGTAGTAAACCAAGAATAGAGAGCTTGGTTGA
ATAGGGCAATGAAGCACGCACACACCGCCCGTCACCCTCCTCAAGTGGCA-CAAGTCAAATATAACCTAT--
TAAAACCAAATAAAACGCAAGAGGAGGCAAGTCGTAACAAGGTAAGCATACTGGAAAGTGTGCTTGGATAAA
C
```

Spurious sequences

☐ Delete spurious sequences before trimming?

Select method

Method: Automated 1

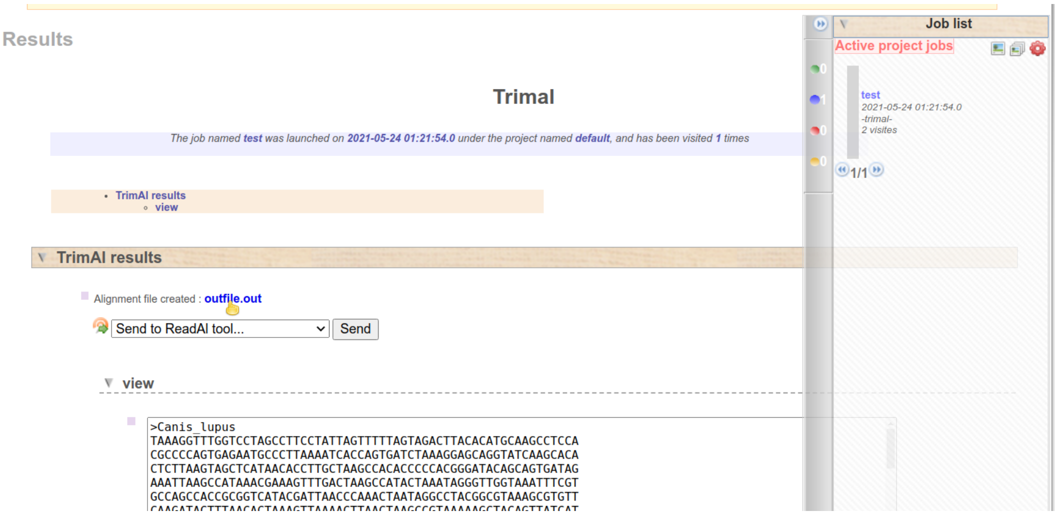
Job

Job name: test

Job description:

Run

1. Paste one of your alignments in the alignment box.
2. Select the trimming algorithm and hit run. The results may take a few minutes to come back.
3. When the job is done, the JOB panel to the right will indicate the job finished and can be viewed.
4. Download or copy paste into Aliview the trimmed file. Do you see the difference? Don't forget to save the trimmed file.



Trimal

The job named **test** was launched on **2021-05-24 01:21:54.0** under the project named **default**, and has been visited **1** times

test
2021-05-24 01:21:54.0
-trimal-
2 visites

TrimAl results

Alignment file created : outfile.out

```
>Canis_lupus
TAAAGGTTTGGTCCTAGCCTTCCTATTAGTTTTTAGTAGACTTACACATGCAAGCCTCCA
CGCCCCAGTGAGAATGCCCTTAAAATCACCAGTGATCTAAAGGAGCAGGTATCAAGCACA
CTCTTAAGTAGCTCATAACACCTTGCTAAGCCACACCCCCACGGGATACAGCAGTGATAG
AAATTAAGCCATAAACGAAAGTTTGACTAAGCCATACTAAATAGGGTTGGTAAATTTCGT
GCCAGCCACCGCGGTCATACGATTAACCCAAACTAATAGGCCTACGGCGTAAAGCGTGTT
CAAGATACTTTAACACTAAAGTTAAAACTTAACTAAGCCGTAAAAAGCTACAGTTATCAT
```

3.3.3 Tree inference using IQTREE in the command prompt

1. Make sure that the IQTREE binary (executable) is in the same folder as the alignments partition.
2. Navigate to that folder in the command prompt
3. To verify IQTREE is properly installed, run:

```
iqtree -h
```

This should print the command line help of the program.

- Using a text editor write a NEXUS format partition definition file as follow:

  ```
  #NEXUS
  begin sets;
  charset cytb = cytb.aln: 1-1140;
  charset r12S = r12S.aln: 1 -985;
  charset fes = fes.aln: 1 - 417;
  charset irbp1st = irbp.aln: 1-1275\3;
  charset irbp2nd = irbp.aln: 2-1275\3;
  charset irbp3rd = irbp.aln: 3-1275\3;
  end;
  ```

- Save the file on the working directory using a meaningful name: e. g. "partitions.nex"
- Run IQTREE with model selection for each partition and 1000 ultra-fast bootstrap resampling.

```
iqtree -spp partitions.nex -m MFP+MERGE -bb 1000 -nt AUTO -pre urs
```

References

[1] Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics 25*, 15 (Aug. 2009), 1972–1973.

[2] Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 5 (Mar. 2004), 1792–1797.

[3] Fulton, T. L., and Strobeck, C. Molecular phylogeny of the Arctoidea (Carnivora): Effect of missing data on supertree and supermatrix analyses of multiple gene data sets. *Molecular Phylogenetics and Evolution 41*, 1 (Oct. 2006), 165–181.

[4] Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution 37*, 5 (May 2020), 1530–1534.

[5] Talavera, G., and Castresana, J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology 56*, 4 (Aug. 2007), 564–577.

[6] Tamura, K., Stecher, G., and Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution*, msab120 (Apr. 2021).