DNA sequences I

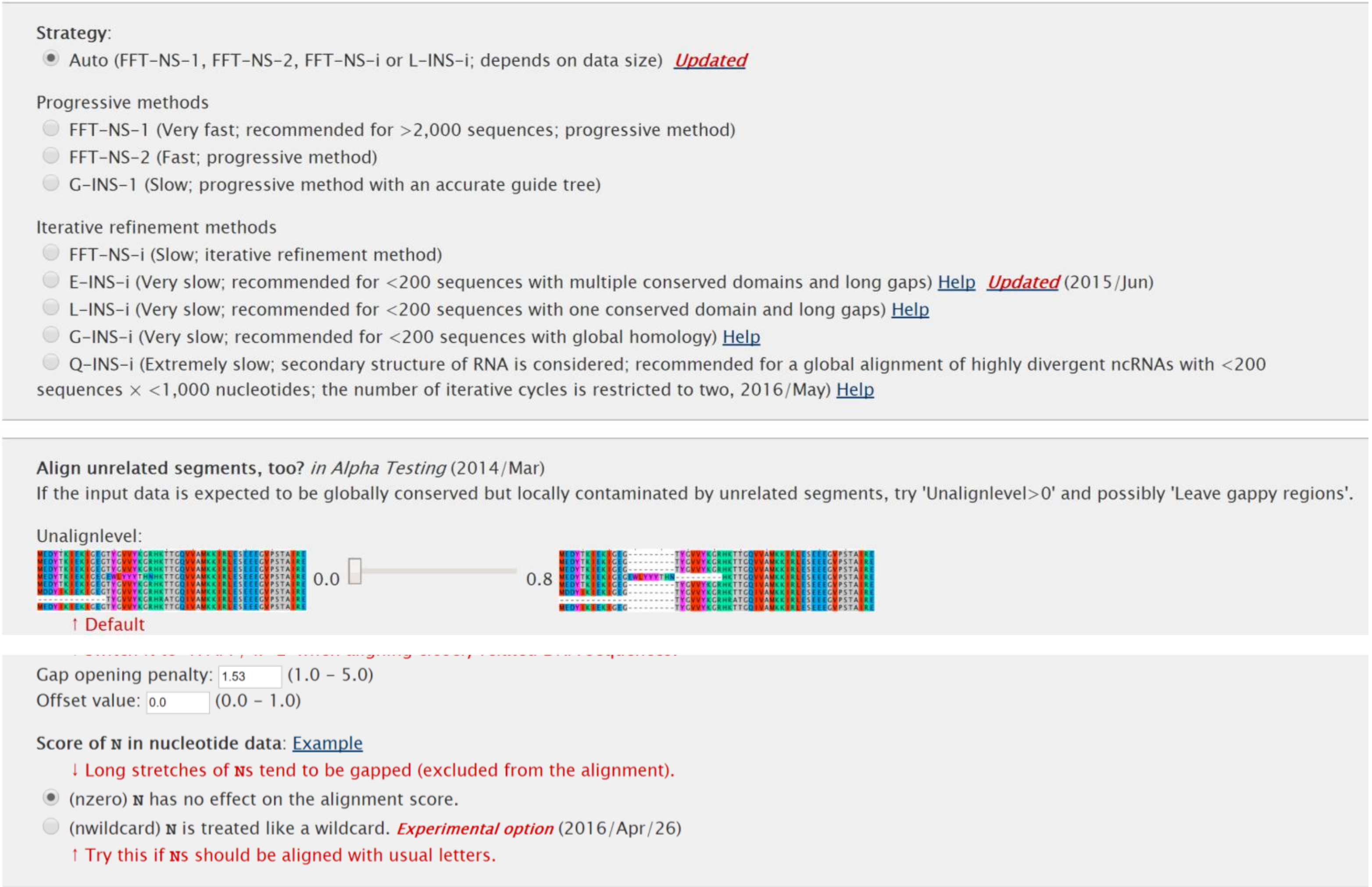
Alignment and Sanger sequence editing, DNA saturation, basic single-gene tree construction

DNA sequences I

- Alignment [Mafft]
- Alignment editing [BioEdit, MEGA]
- Alignment improvement [Gblocks, Trimal]
- Alignment conversion, concatenation [FASconCAT]
- Model selection [jModeltest, PartitionFinder]
- Tree reconstruction [RAxML, PAUP, MrBayes]
- Tree manipulation [FigTree, R]

Sequence alignment

- http://mafft.cbrc.jp/alignment/server/



Strategy:

● Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) *Updated*

Progressive methods

○ FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)

○ FFT-NS-2 (Fast; progressive method)

○ G-INS-1 (Slow; progressive method with an accurate guide tree)

Iterative refinement methods

○ FFT-NS-i (Slow; iterative refinement method)

○ E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps) Help *Updated* (2015/Jun)

○ L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps) Help

○ G-INS-i (Very slow; recommended for <200 sequences with global homology) Help

○ Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of highly divergent ncRNAs with <200 sequences × <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) Help

Align unrelated segments, too? *in Alpha Testing* (2014/Mar)

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unalignlevel>0' and possibly 'Leave gappy regions'.

Unalignlevel:

↑ Default

Gap opening penalty: 1.53 (1.0 – 5.0)

Offset value: 0.0 (0.0 – 1.0)

Score of N in nucleotide data: Example

↓ Long stretches of Ns tend to be gapped (excluded from the alignment).

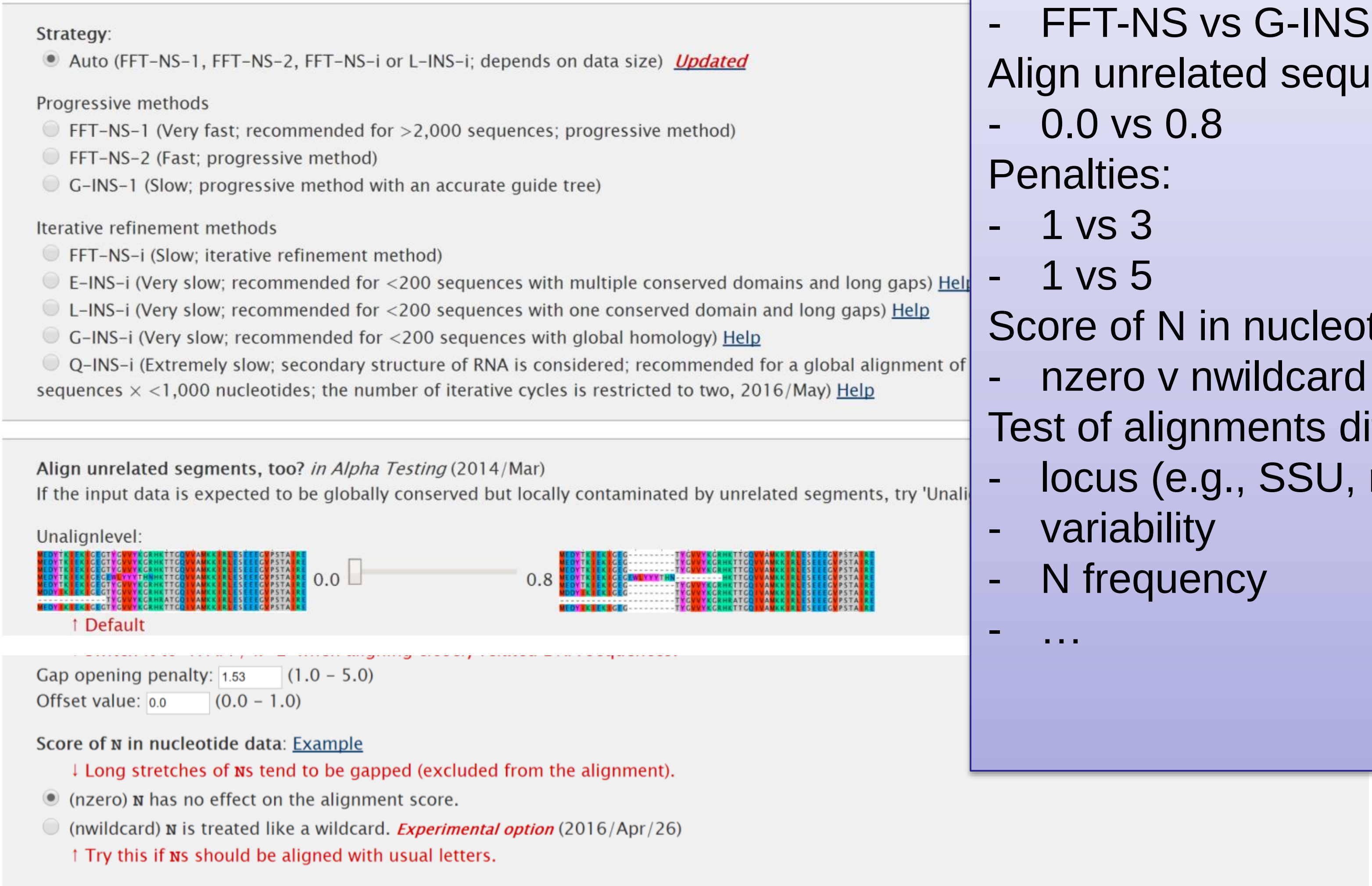
● (nzero) N has no effect on the alignment score.

○ (nwildcard) N is treated like a wildcard. *Experimental option* (2016/Apr/26)

↑ Try this if Ns should be aligned with usual letters.

Sequence alignment

- *Playing with MAFFT options*



Strategy:

● Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) *Updated*

Progressive methods

○ FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)

○ FFT-NS-2 (Fast; progressive method)

○ G-INS-1 (Slow; progressive method with an accurate guide tree)

Iterative refinement methods

○ FFT-NS-i (Slow; iterative refinement method)

○ E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps) Help

○ L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps) Help

○ G-INS-i (Very slow; recommended for <200 sequences with global homology) Help

○ Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of sequences × <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) Help

Align unrelated segments, too? *in Alpha Testing* (2014/Mar)

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unali

Unalignlevel: 0.0 0.8

↑ Default

Gap opening penalty: 1.53 (1.0 – 5.0)

Offset value: 0.0 (0.0 – 1.0)

Score of N in nucleotide data: Example

↓ Long stretches of Ns tend to be gapped (excluded from the alignment).

● (nzero) N has no effect on the alignment score.

○ (nwildcard) N is treated like a wildcard. *Experimental option* (2016/Apr/26)

↑ Try this if Ns should be aligned with usual letters.

Strategy:
- FFT-NS vs Q-INS
- FFT-NS vs G-INS

Align unrelated sequences:
- 0.0 vs 0.8

Penalties:
- 1 vs 3
- 1 vs 5

Score of N in nucleotide data:
- nzero v nwildcard

Test of alignments differing by:
- locus (e.g., SSU, rbcL)
- variability
- N frequency
- …

Sequence editing

- BioEdit - http://www.mbio.ncsu.edu/bioedit/bioedit.html

- A few tips:
  - Sequence – Manipulations – UPPERCASE
  - Sequence – Nucleic Acid – RNA->DNA
  - Alignment – Minimize alignment to mask
  - Edit – Copy sequence titles
  - Search options….

Sequence editing

- MEGA - http://www.megasoftware.net/

MEGA Molecular Evolutionary Genetics Analysis

- A few tips (.fas):
  - Data – Reverse Complement
  - Data – Translate/untranslate
- A few tips (.meg):
  - Distance – Compute pairwise distance
  - Phylogeny – Construct fast trees
  - Statistics (Data explorer) – Nucleotide composition
  - Highlight (Data explorer) – Mark sites

Sequence editing

- FaBox- http://users-birc.au.dk/biopv/php/fabox/

- Edit names
- Crop and merge alignments
- Extract variable sites („Show variable sites only“)
- Convert fasta to other formats (e.g., TCS input)
- Create MrBayes nexus file („Create MrBayes input file from fasta (fasta2mrbayes)“

Sequence editing

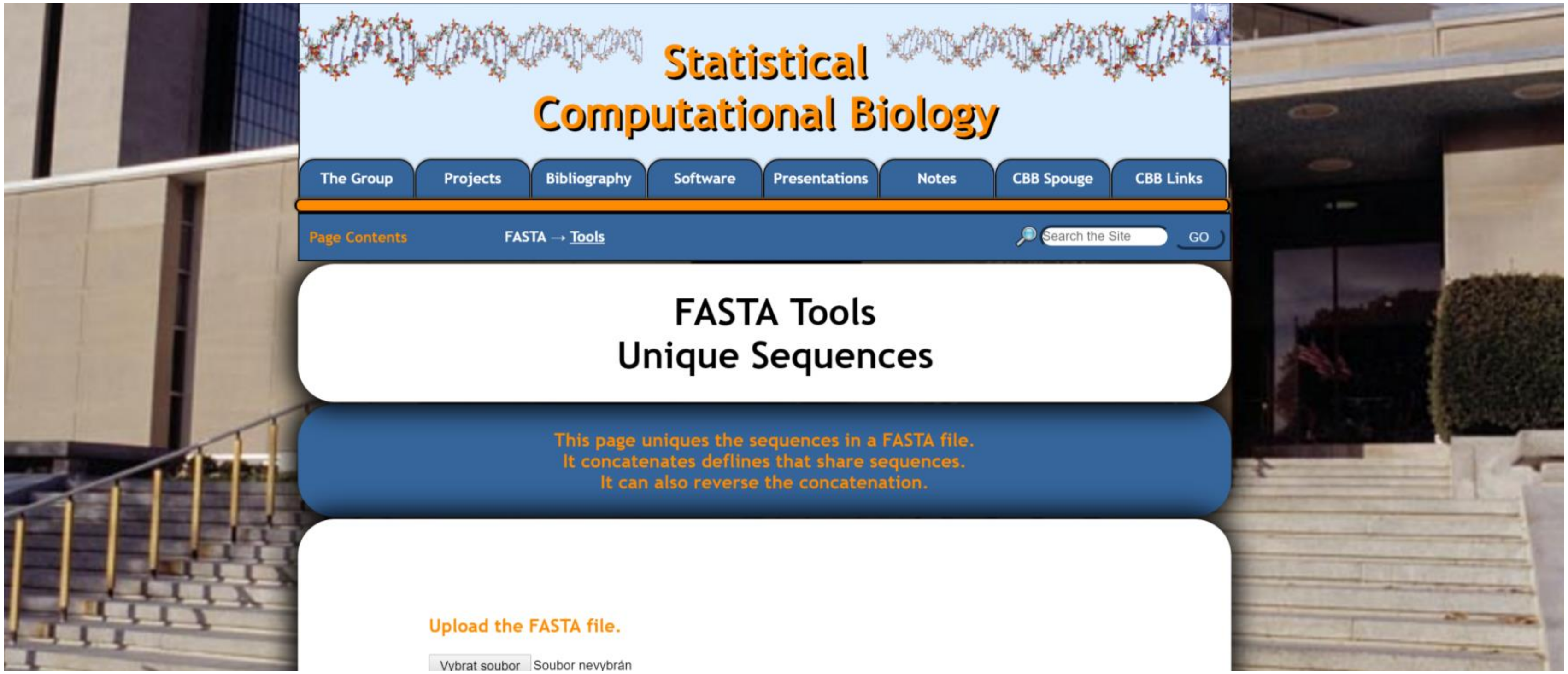
- SeqState - http://bioinfweb.info/Software/SeqState

- Coding gaps as a special state

- *File -> Load NEXUS file*
- *IndelCoder -> simple indel coding*

Sequence editing

- FASTA Tools – Unique sequences
- https://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/fasta/uniqueseq.html



FASTA Tools
Unique Sequences

This page uniques the sequences in a FASTA file.
It concatenates deflines that share sequences.
It can also reverse the concatenation.

Upload the FASTA file.

Vybrat soubor Soubor nevybrán

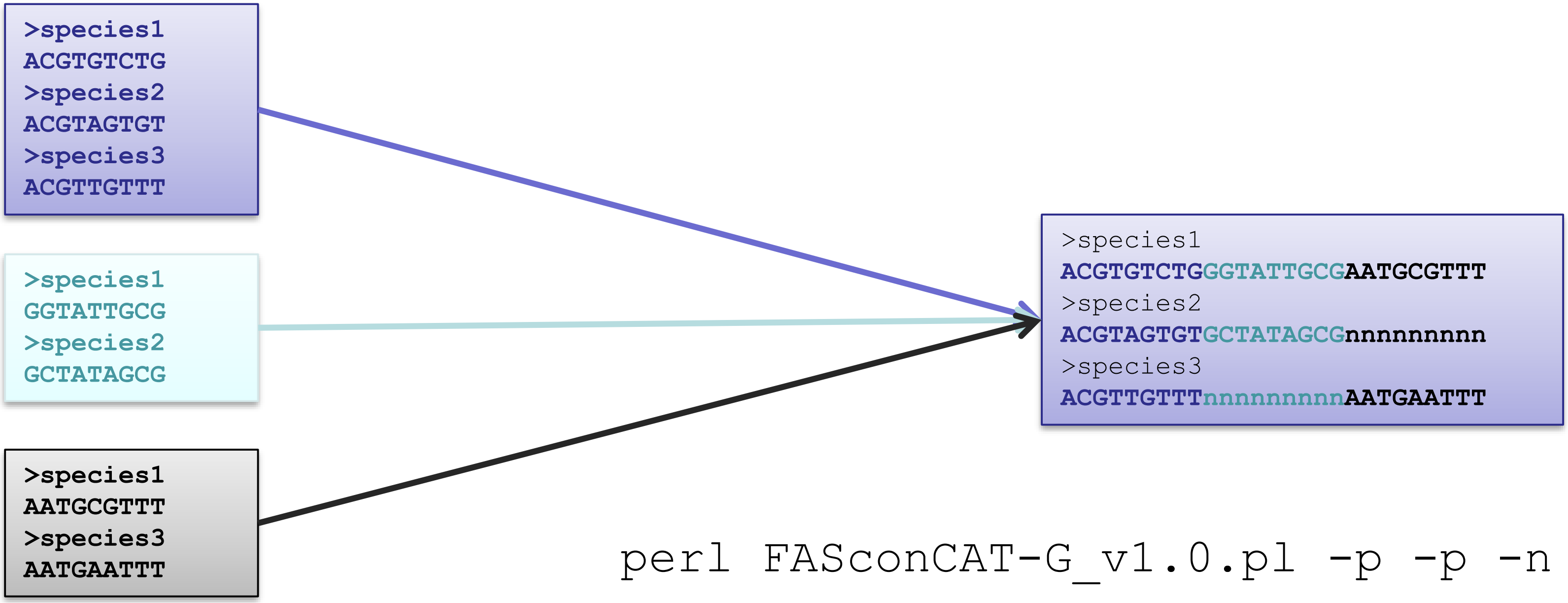
Sequence editing

- A few tips in R:
  - Save sequence names
  - Replace sequence names
  - Minimize alignment to mask
  - Save an alignment of selected sequences
  - Concatenate alignments

Alignment concatenations, conversion

FASconCAT – https://www.zfmk.de/en/research/research-centres-and-groups/fasconcat-g

- Perl script
- concatenating alignments (with same headers but not necessarily with all samples in all alignments)
- conversion between fasta, phylip and nexus



```
>species1
ACGTGTCTG
>species2
ACGTAGTGT
>species3
ACGTTGTTT
```

```
>species1
GGTATTGCG
>species2
GCTATAGCG
```

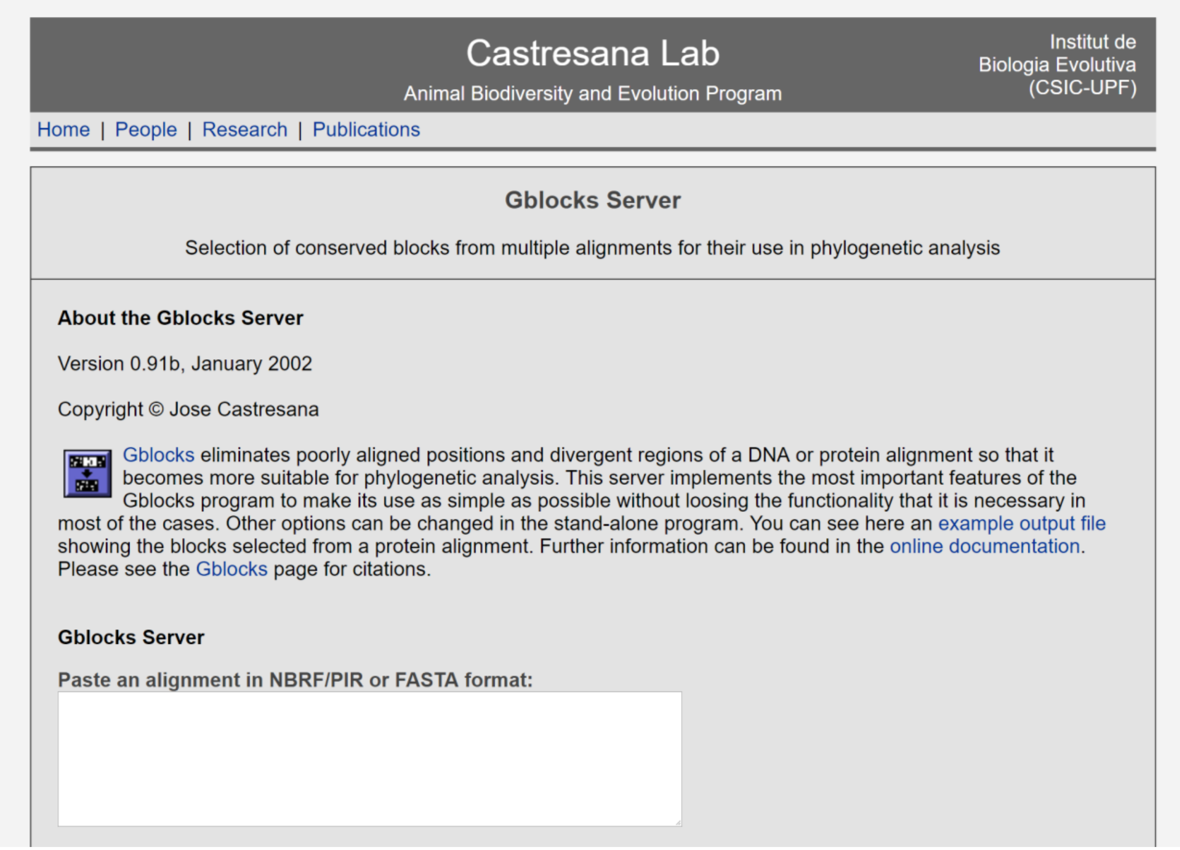
```
>species1
AATGCGTTT
>species3
AATGAATTT
```

```
>species1
ACGTGTCTGGGTATTGCGAATGCGTTT
>species2
ACGTAGTGTGCTATAGCGnnnnnnnnn
>species3
ACGTTGTTTnnnnnnnnnAATGAATTT
```

```
perl FASconCAT-G_v1.0.pl -p -p -n -n -s
```

Alignment improvement

- Gblocks:
- http://molevol.cmima.csic.es/castresana/Gblocks_server.html



Castresana Lab

Animal Biodiversity and Evolution Program

Institut de
Biologia Evolutiva
(CSIC-UPF)

Home | People | Research | Publications

Gblocks Server

Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis

About the Gblocks Server

Version 0.91b, January 2002

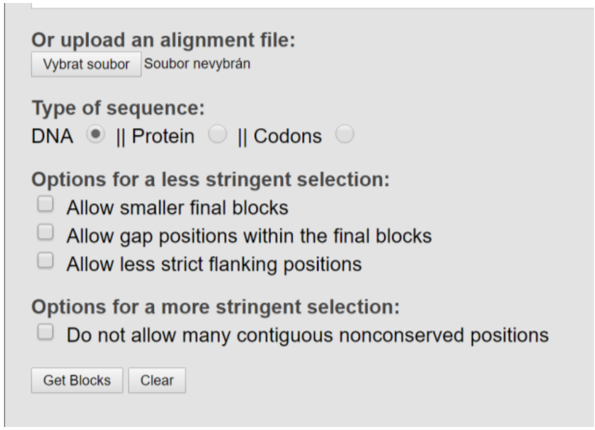
Copyright © Jose Castresana

Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment so that it becomes more suitable for phylogenetic analysis. This server implements the most important features of the Gblocks program to make its use as simple as possible without loosing the functionality that it is necessary in most of the cases. Other options can be changed in the stand-alone program. You can see here an example output file showing the blocks selected from a protein alignment. Further information can be found in the online documentation. Please see the Gblocks page for citations.

Gblocks Server

Paste an alignment in NBRF/PIR or FASTA format:

Options to play with:



Or upload an alignment file:

Vybrat soubor Soubor nevybrán

Type of sequence:

DNA ● || Protein ○ || Codons ○

Options for a less stringent selection:

□ Allow smaller final blocks

□ Allow gap positions within the final blocks

□ Allow less strict flanking positions

Options for a more stringent selection:

□ Do not allow many contiguous nonconserved positions

Alignment improvement

- Gblocks

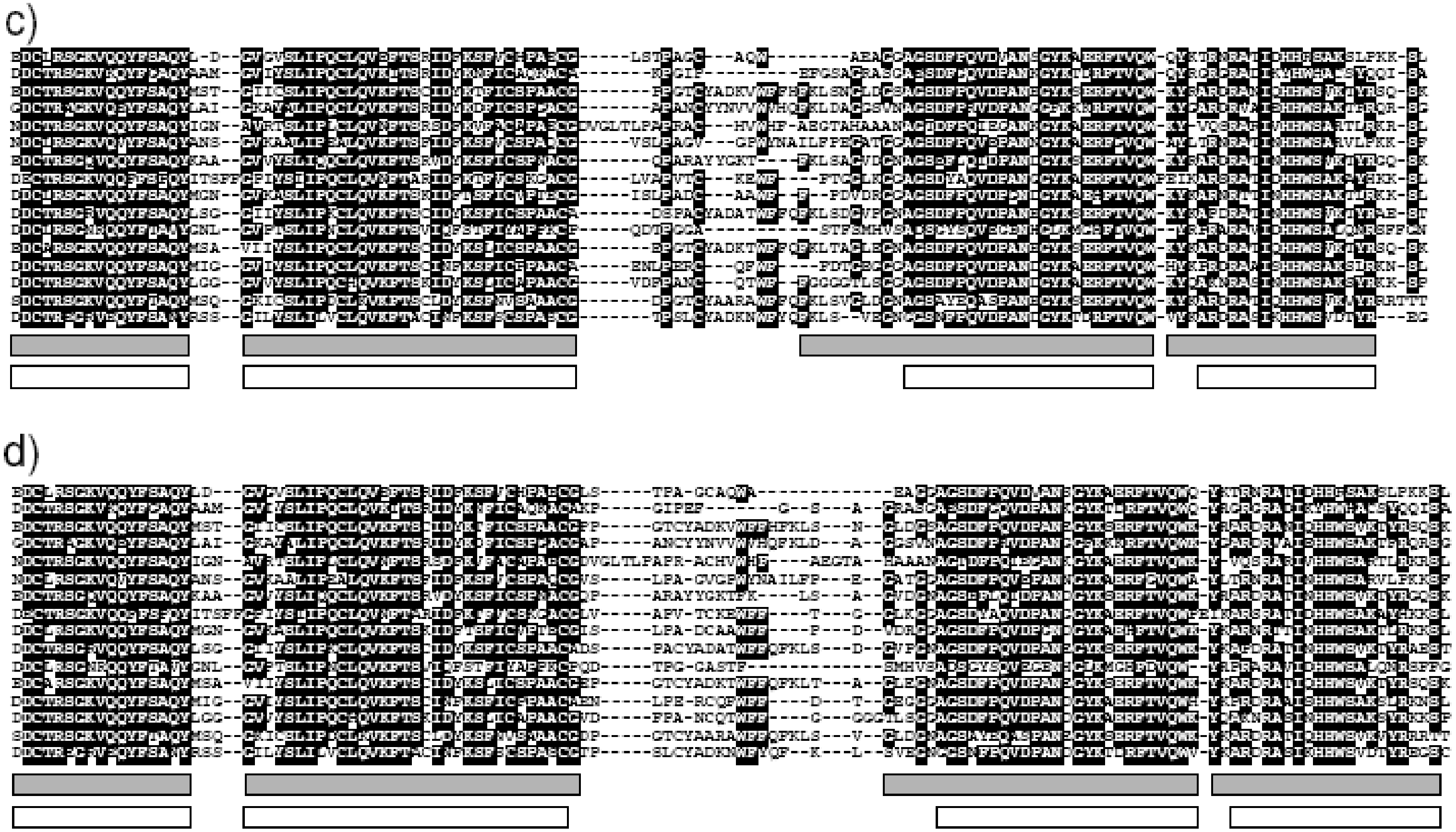
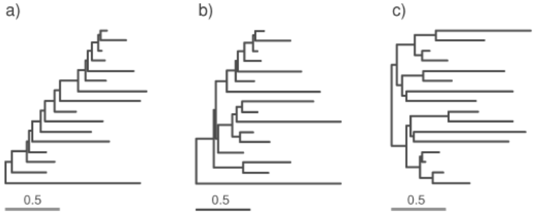
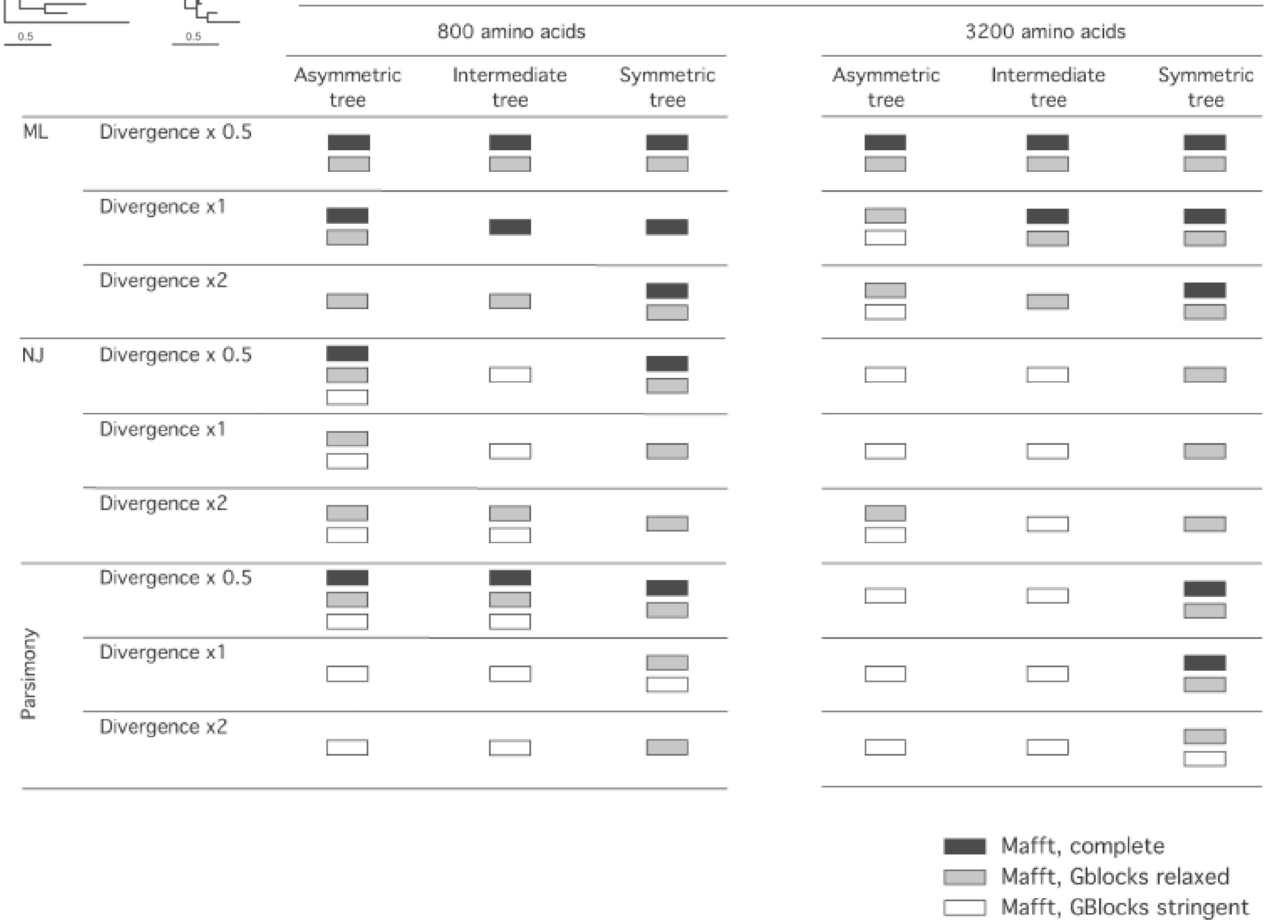


FIGURE 2. Fragment of a simulated alignment (a) and the realignment of the same sequences (after gap removal) by ClustalW (b), Mafft (c), and Probcons (d). The simulation corresponds to an asymmetric tree with divergence ×1. The blocks below each alignment represent the fragments selected by Gblocks with relaxed conditions (grey blocks) and with stringent conditions (white blocks). Positions of the alignments where more than 50% of the sequences are identical are shown with black boxes.

Alignment improvement





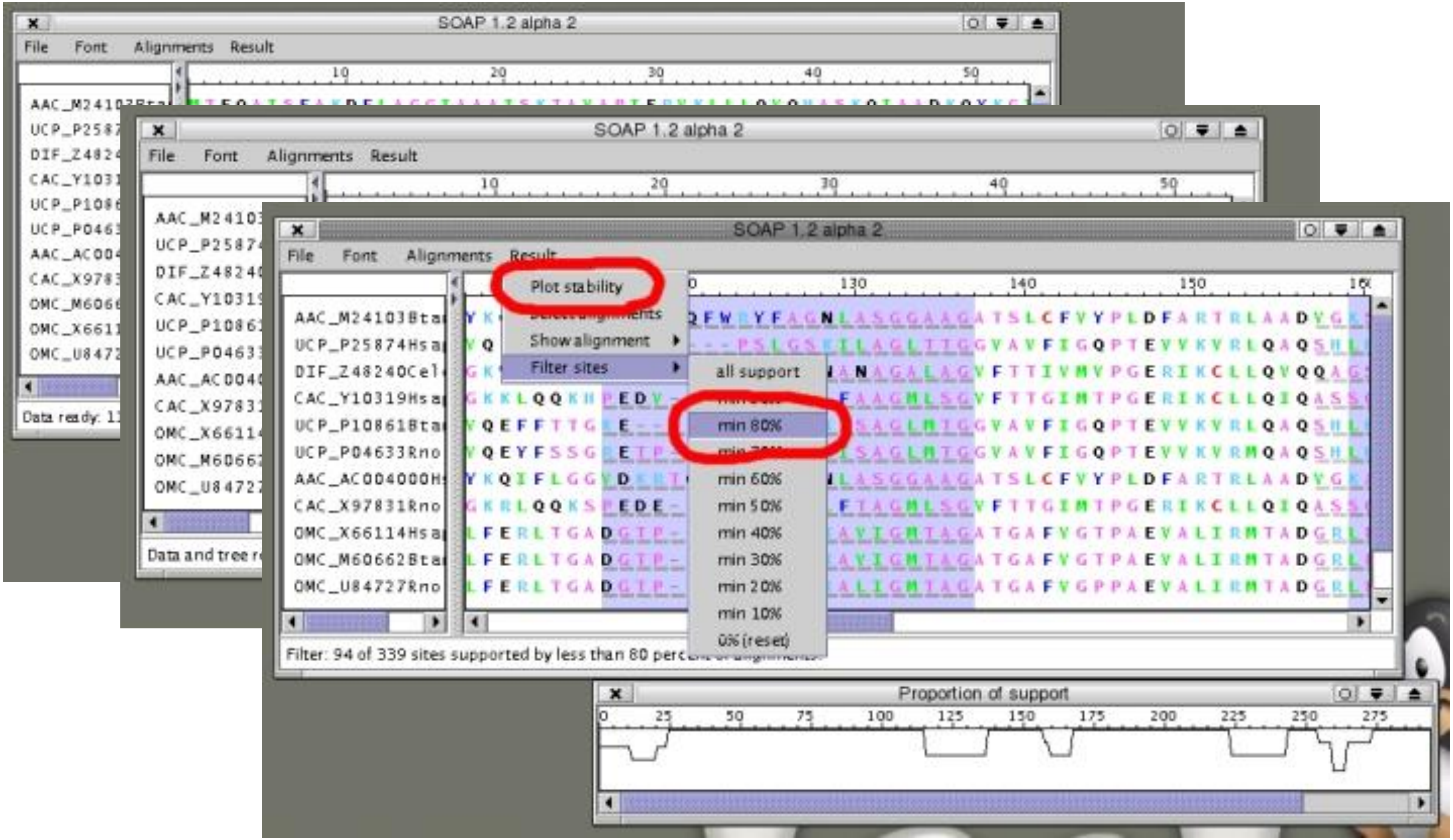
		800 amino acids			3200 amino acids		
		Asymmetric tree	Intermediate tree	Symmetric tree	Asymmetric tree	Intermediate tree	Symmetric tree
ML	Divergence x 0.5	■ ▨	■ ▨	■ ▨	■ ▨	■ ▨	■ ▨
	Divergence x1	■ ▨	■	■	▨ □	■ ▨	■ ▨
	Divergence x2	▨	▨	■ ▨	▨ □	▨	■ ▨
NJ	Divergence x 0.5	■ ▨ □	□	■ ▨	□	□	▨
	Divergence x1	▨ □	□	▨	□	□	▨
	Divergence x2	▨ □	▨ □	▨	▨ □	□	▨
Parsimony	Divergence x 0.5	■ ▨ □	■ ▨ □	■ ▨	□	□	■ ▨
	Divergence x1	□	□	▨ □	□	□	■ ▨
	Divergence x2	□	□	▨	□	□	▨ □

■ Mafft, complete
▨ Mafft, Gblocks relaxed
□ Mafft, GBlocks stringent

FIGURE 6. Mafft alignment strategies that give rise to the statistically best topologies. When two or more strategies do not show statistical differences in Robinson-Foulds distances, all equivalent strategies are represented. The complete alignment is represented by a black block, and the relaxed and stringent Gblocks strategies by grey and white blocks, respectively.

Alignment improvement

- SOAP - http://ueg.ulb.ac.be/SOAP/



Data ready: 11

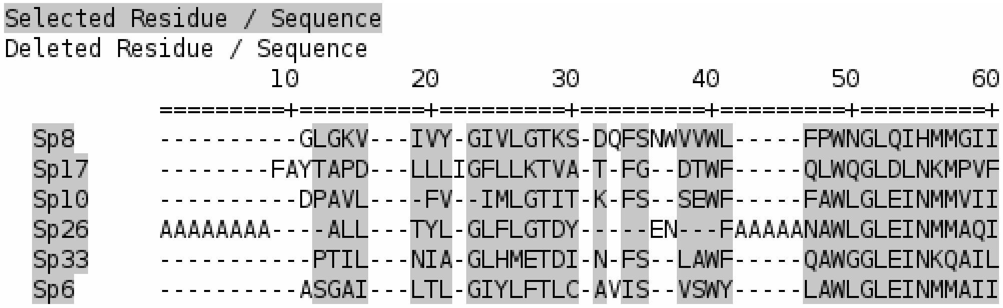
Data and tree r

Filter: 94 of 339 sites supported by less than 80 percent of alignments.

Alignment improvement

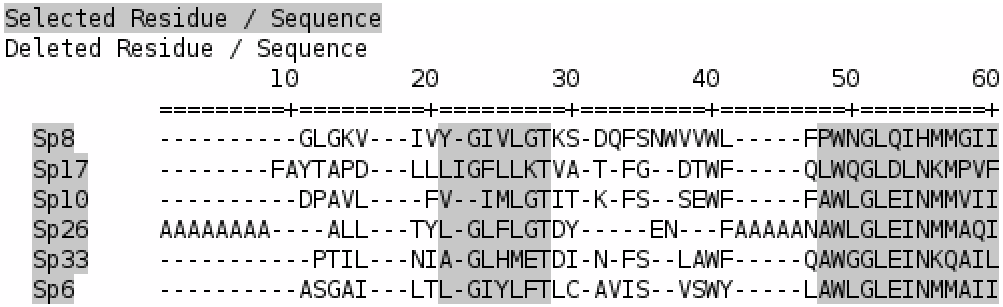
- Trimal - http://trimal.cgenomics.org
  - automated removal of spurious sequences or poorly aligned regions

```
trimal -in example1 -out output6 -htmlout output6.html -gappyout
```



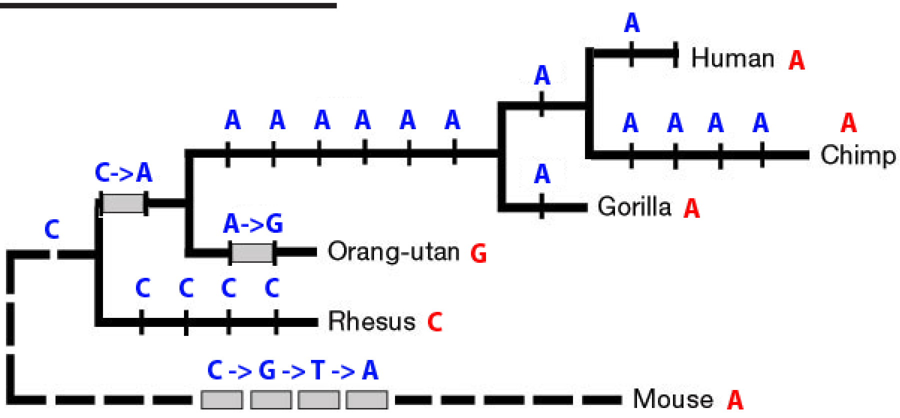
```
Selected Residue / Sequence
Deleted Residue / Sequence
                    10        20        30        40        50        60
            =========+=========+=========+=========+=========+=========+
  Sp8       ----------GLGKV---IVY-GIVLGTKS-DQFSNWVVWL-----FPWNGLQIHMMGII
  Sp17      --------FAYTAPD---LLLIGFLLKTVA-T-FG--DTWF-----QLWQGLDLNKMPVF
  Sp10      ----------DPAVL----FV--IMLGTIT-K-FS--SEWF-----FAWLGLEINMMVII
  Sp26      AAAAAAAA----ALL---TYL-GLFLGTDY-----EN---FAAAAANAWLGLEINMMAQI
  Sp33      -----------PTIL---NIA-GLHMETDI-N-FS--LAWF-----QAWGGLEINKQAIL
  Sp6       ----------ASGAI---LTL-GIYLFTLC-AVIS--VSWY-----LAWLGLEINMMAII
```

```
trimal -in example1 -out output7 -htmlout output7.html -strict
```



```
Selected Residue / Sequence
Deleted Residue / Sequence
                    10        20        30        40        50        60
            =========+=========+=========+=========+=========+=========+
  Sp8       ----------GLGKV---IVY-GIVLGTKS-DQFSNWVVWL-----FPWNGLQIHMMGII
  Sp17      --------FAYTAPD---LLLIGFLLKTVA-T-FG--DTWF-----QLWQGLDLNKMPVF
  Sp10      ----------DPAVL----FV--IMLGTIT-K-FS--SEWF-----FAWLGLEINMMVII
  Sp26      AAAAAAAA----ALL---TYL-GLFLGTDY-----EN---FAAAAANAWLGLEINMMAQI
  Sp33      -----------PTIL---NIA-GLHMETDI-N-FS--LAWF-----QAWGGLEINKQAIL
  Sp6       ----------ASGAI---LTL-GIYLFTLC-AVIS--VSWY-----LAWLGLEINMMAII
```

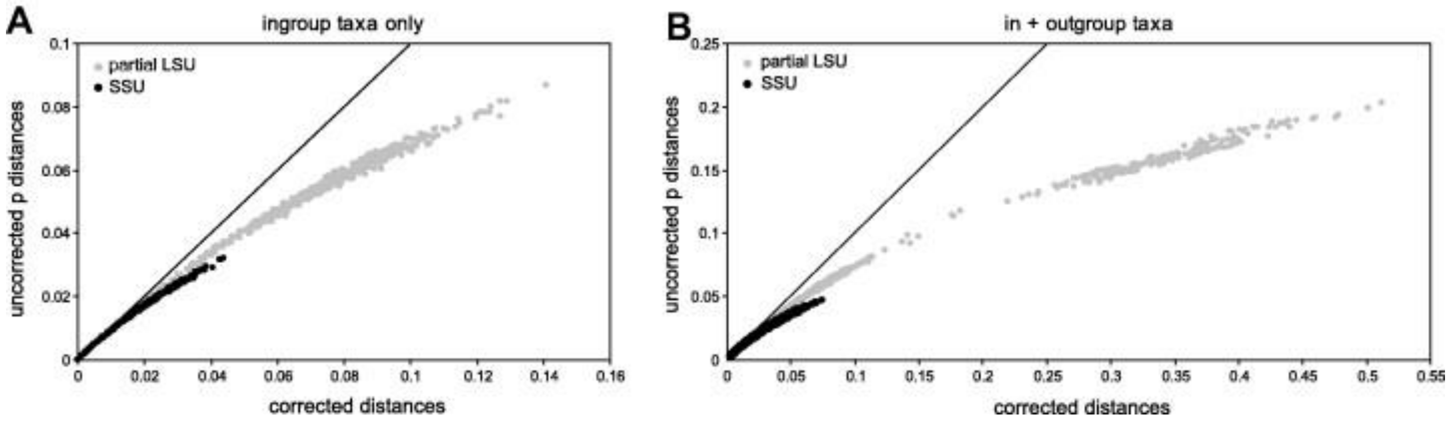
Alignment improvement



Substitution saturation

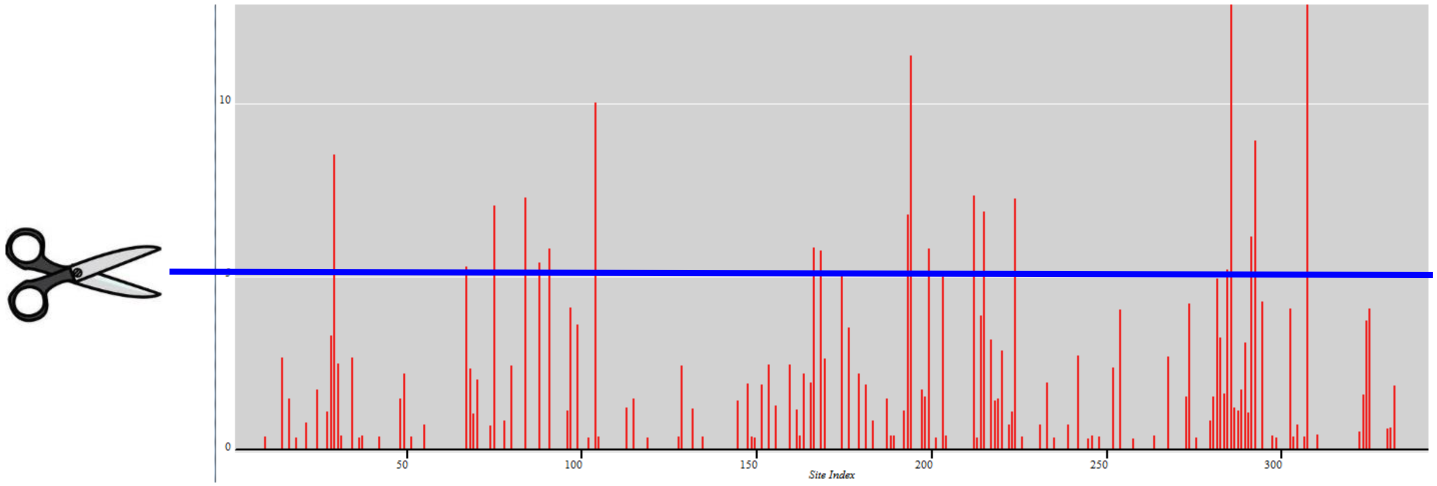
- Some positions in alignment were changed multiple times
- Due to only 4 nucleotide types, the noise is stochastically incresing by time
- Saturated positions can represent the majority of variability in data
- A big problem for MP analyses!

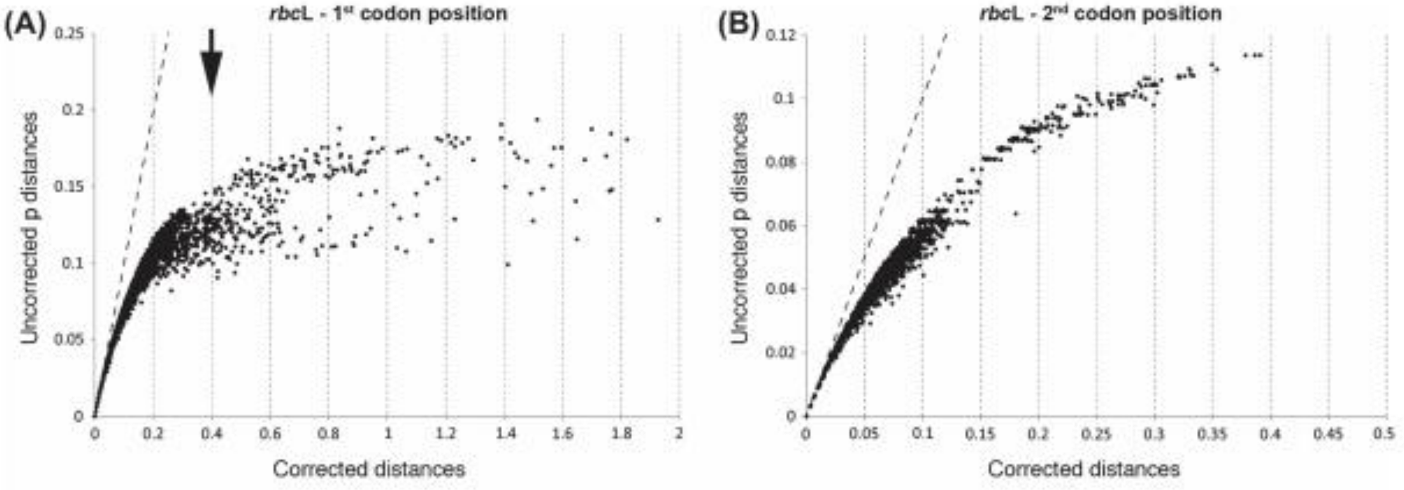
1) Saturation curves:
   - Comparison of sequence distances accounted under the simple and more complex substitution models



Alignment improvement

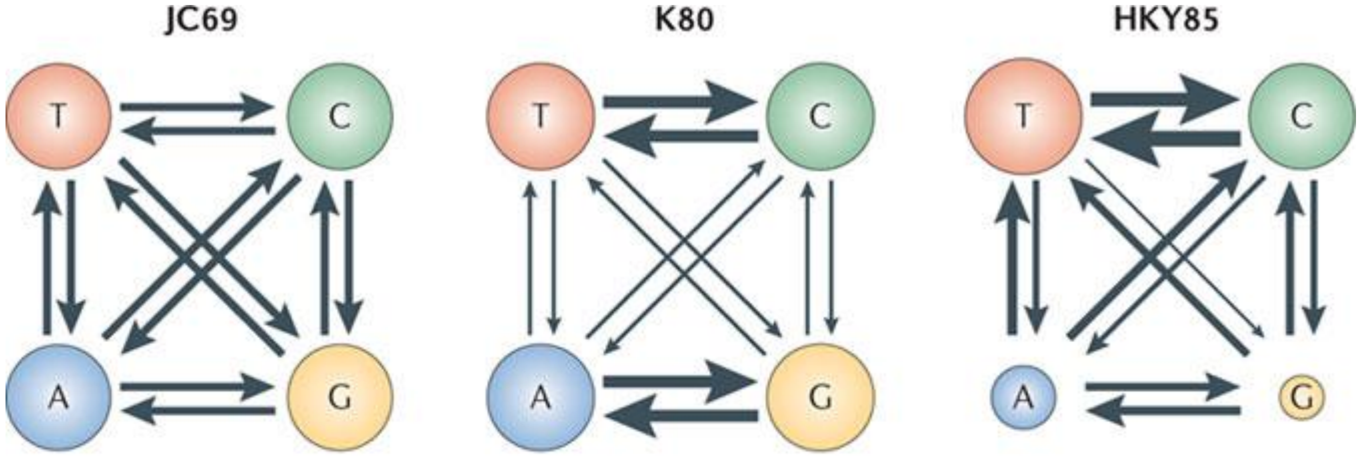
2) Site stripping
   - Deletion of saturated alignment positions





Model selection

- *p-distance*: the simplest distance measure: the proportion of sites that differ between two sequences (saturation!).
- *substitution models* – models of how frequently different mutations occur in the DNA to precisely estimate the evolutionary distance between organisms
- Different combinations of base frequencies and nucleotide substitutions



Nature Reviews | Genetics

Model selection

- Jukes-Cantor (JC69): equal base frequencies, all substitutions equally likely
(nst=1, rate classification: aaaaaa)

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

- Felsenstein (F81): variable base frequencies, all substitutions equally likely
(nst=1, rate classification: aaaaaa)

	A	T	C	G
A	-	$\alpha \Pi_T$	$\alpha \Pi_C$	$\alpha \Pi_G$
T	$\alpha \Pi_A$	-	$\alpha \Pi_C$	$\alpha \Pi_G$
C	$\alpha \Pi_A$	$\alpha \Pi_T$	-	$\alpha \Pi_G$
G	$\alpha \Pi_A$	$\alpha \Pi_T$	$\alpha \Pi_C$	-

- Kimura (K80): equal base frequencies, one transition rate and one transversion rate
(nst=2, rate classification: abaaba)

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

- Hasegawa-Kishino-Yano (HKY): variable base frequencies, one transition rate and one transversion rate
(nst=2, rate classification: abbbba)

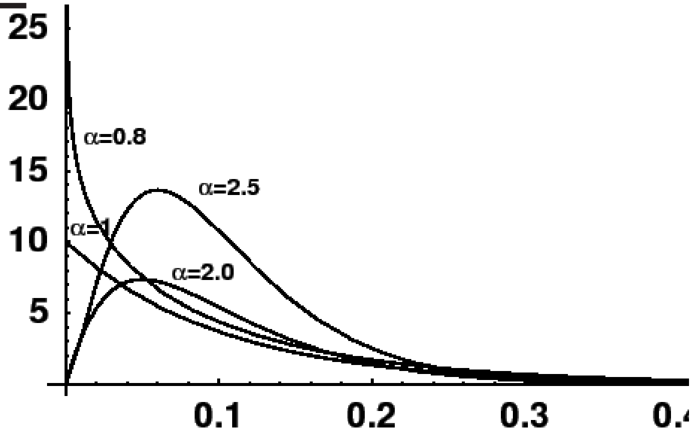
	A	T	C	G
A	-	$\beta \Pi_T$	$\beta \Pi_C$	$\alpha \Pi_G$
T	$\beta \Pi_A$	-	$\beta \Pi_C$	$\beta \Pi_G$
C	$\beta \Pi_A$	$\beta \Pi_T$	-	$\beta \Pi_G$
G	$\beta \Pi_A$	$\beta \Pi_T$	$\beta \Pi_C$	-

- General time reverdible (GTRvariable base frequencies, symmetrical substitution matrix
(nst=6, rate classification: abcdef)

	A	T	C	G
A	-	$\alpha \Pi_T$	$\beta \Pi_C$	$\gamma \Pi_G$
T	$\alpha \Pi_A$	-	$\delta \Pi_C$	$\varepsilon \Pi_G$
C	$\beta \Pi_A$	$\delta \Pi_T$	-	$\zeta \Pi_G$
G	$\alpha \Pi_A$	$\varepsilon \Pi_T$	$\zeta \Pi_C$	-

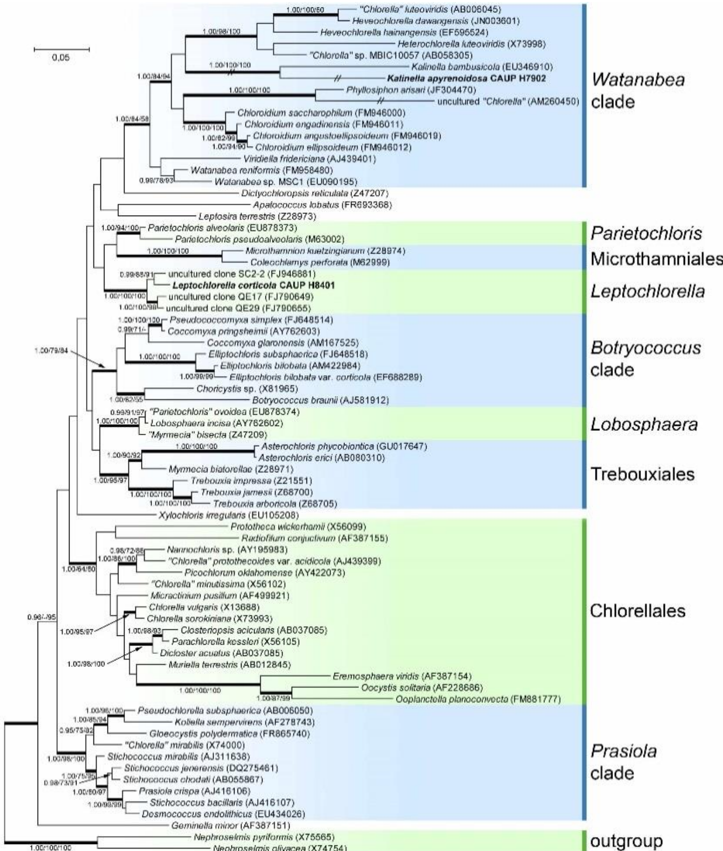
Model selection

- Gamma distribution (Γ): models a variability in substitution rates on different alignment positions. Usually, a model is simplified to 4 α categories



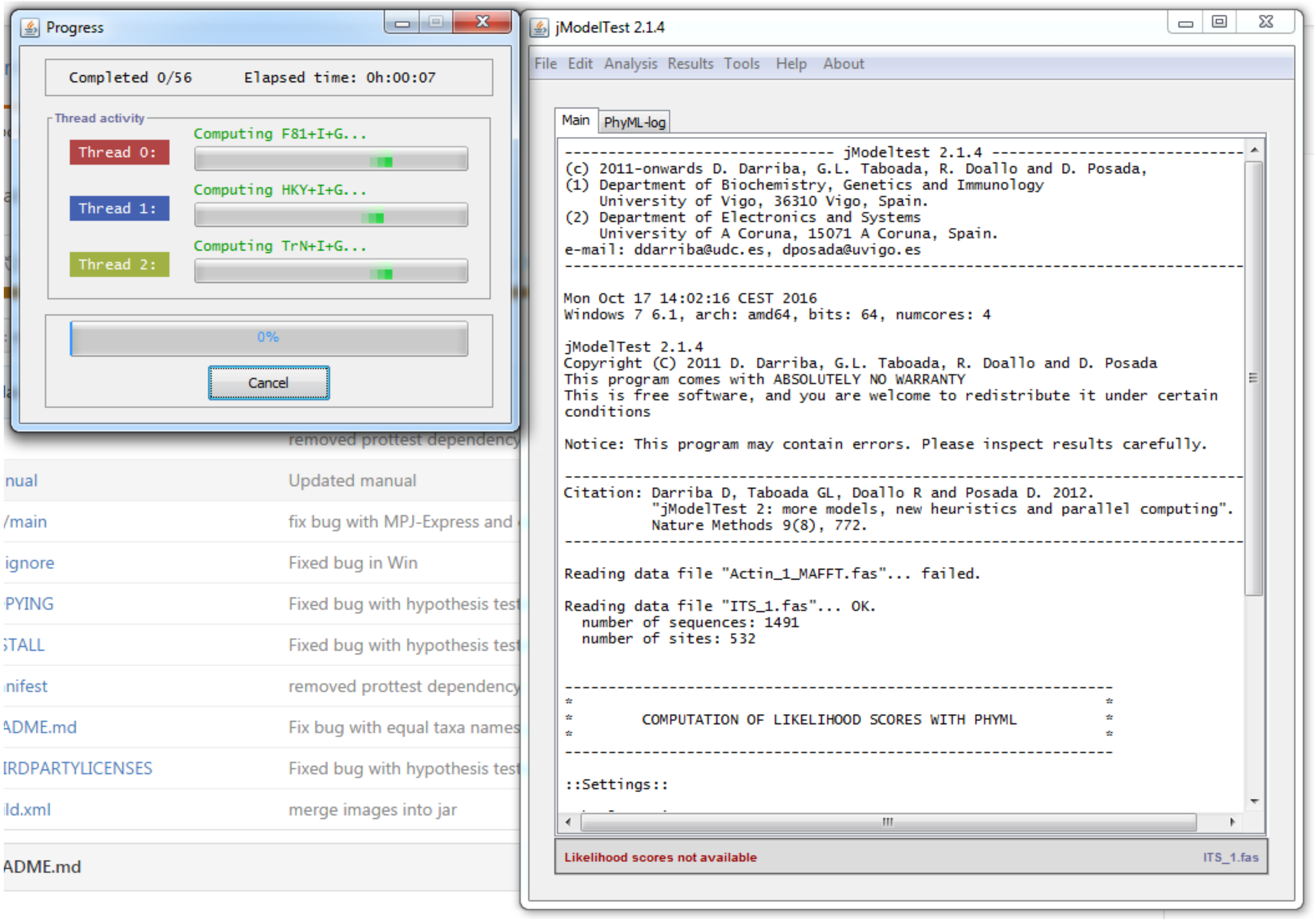
- Proportion of invariable sites (I): existence of a majority of invariable sites may negatively affect the estimated genetic distances. I model is particularly important in joint occurrence of very short and long branches

- Covarion (cov): models a variability in nucleotide substitution rates across the phylogenetic tree


0,05
1.00/100/60
"Chlorella" luteoviridis (AB006045)
Heveochlorella dawangensis (JN003601)
1.00/98/100
Heveochlorella hainangensis (EF595524)
Heterochlorella luteoviridis (X73998)
"Chlorella" sp. MBIC10057 (AB058305)
1.00/100/100
Kalinella bambusicola (EU346910)
1.00/94/94
Kalinella apyrenoidosa CAUP H7902
1.00/100/100
Phyllosiphon arisari (JF304470)
uncultured "Chlorella" (AM260450)
1.00/84/58
Chloroidium saccharophilum (FM946000)
1.00/100/100
Chloroidium engadinensis (FM946011)
1.00/82/99
Chloroidium angustoellipsoideum (FM946019)
1.00/94/90
Chloroidium ellipsoideum (FM946012)
Viridiella fridericiana (AJ439401)
Watanabea reniformis (FM958480)
0.99/78/93
Watanabea sp. MSC1 (EU090195)
Dictyochloropsis reticulata (Z47207)
Apatococcus lobatus (FR693368)
Leptosira terrestris (Z28973)
Watanabea clade
1.00/94/100
Parietochloris alveolaris (EU878373)
Parietochloris pseudoalveolaris (M63002)
Parietochloris
1.00/100/100
Microthamnion kuetzingianum (Z28974)
Coleochlamys perforata (M62999)
Microthamniales
0.99/88/91
uncultured clone SC2-2 (FJ946881)
Leptochlorella corticola CAUP H8401
1.00/100/100
uncultured clone QE17 (FJ790649)
1.00/100/98
uncultured clone QE29 (FJ790655)
Leptochlorella
1.00/100/100
Pseudococcomyxa simplex (FJ648514)
0.99/71/-
Coccomyxa pringsheimii (AY762603)
Coccomyxa glaronensis (AM167525)
1.00/79/84
1.00/100/100
Elliptochloris subsphaerica (FJ648518)
Elliptochloris bilobata (AM422984)
1.00/99/99
Elliptochloris bilobata var. corticola (EF688289)
Choricystis sp. (X81965)
1.00/82/55
Botryococcus braunii (AJ581912)
Botryococcus clade
0.99/91/97
"Parietochloris" ovoidea (EU878374)
1.00/100/100
Lobosphaera incisa (AY762602)
"Myrmecia" bisecta (Z47209)
Lobosphaera
1.00/100/100
Asterochloris phycobiontica (GU017647)
1.00/90/92
Asterochloris erici (AB080310)
Myrmecia biatorellae (Z28971)
1.00/95/97
Trebouxia impressa (Z21551)
1.00/100/100
Trebouxia jamesii (Z68700)
1.00/100/100
Trebouxia arboricola (Z68705)
Trebouxiales
Xylochloris irregularis (EU105208)
Prototheca wickerhamii (X56099)
Radiofilum conjuctivum (AF387155)
0.98/72/88
Nannochloris sp. (AY195983)
1.00/86/100
"Chlorella" protothecoides var. acidicola (AJ439399)
1.00/64/80
Picochlorum oklahomense (AY422073)
"Chlorella" minutissima (X56102)
Micractinium pusillum (AF499921)
Chlorella vulgaris (X13688)
0.96/-/95
Chlorella sorokiniana (X73993)
1.00/95/97
1.00/98/93
Closteriopsis acicularis (AB037085)
Parachlorella kessleri (X56105)
Dicloster acuatus (AB037085)
1.00/98/100
Muriella terrestris (AB012845)
Eremosphaera viridis (AF387154)
1.00/100/100
Oocystis solitaria (AF228686)
1.00/87/99
Ooplanctella planoconvecta (FM881777)
Chlorellales
1.00/96/100
Pseudochlorella subsphaerica (AB006050)
1.00/85/94
Koliella sempervirens (AF278743)
0.95/75/82
Gloeocystis polydermatica (FR865740)
"Chlorella" mirabilis (X74000)
1.00/98/100
Stichococcus mirabilis (AJ311638)
1.00/75/95
Stichococcus jenerensis (DQ275461)
0.98/73/91
Stichococcus chodati (AB055867)
1.00/80/97
Prasiola crispa (AJ416106)
1.00/98/99
Stichococcus bacillaris (AJ416107)
Desmococcus endolithicus (EU434026)
Prasiola clade
Geminella minor (AF387151)
Nephroselmis pyriformis (X75565)
1.00/100/100
Nephroselmis olivacea (X74754)
outgroup

Model selection

- jModelTest: https://github.com/ddarriba/jmodeltest2



Completed 0/56 Elapsed time: 0h:00:07

Computing F81+I+G...

Computing HKY+I+G...

Computing TrN+I+G...

	removed prottest dependency
nual	Updated manual
/main	fix bug with MPJ-Express and
ignore	Fixed bug in Win
PYING	Fixed bug with hypothesis test
STALL	Fixed bug with hypothesis test
nifest	removed prottest dependency
ADME.md	Fix bug with equal taxa names
IRDPARTYLICENSES	Fixed bug with hypothesis test
ld.xml	merge images into jar

ADME.md

```
------------------------------- jModeltest 2.1.4 ------------------------------
(c) 2011-onwards D. Darriba, G.L. Taboada, R. Doallo and D. Posada,
(1) Department of Biochemistry, Genetics and Immunology
    University of Vigo, 36310 Vigo, Spain.
(2) Department of Electronics and Systems
    University of A Coruna, 15071 A Coruna, Spain.
e-mail: ddarriba@udc.es, dposada@uvigo.es
--------------------------------------------------------------------------------

Mon Oct 17 14:02:16 CEST 2016
Windows 7 6.1, arch: amd64, bits: 64, numcores: 4

jModelTest 2.1.4
Copyright (C) 2011 D. Darriba, G.L. Taboada, R. Doallo and D. Posada
This program comes with ABSOLUTELY NO WARRANTY
This is free software, and you are welcome to redistribute it under certain
conditions

Notice: This program may contain errors. Please inspect results carefully.

--------------------------------------------------------------------------------
Citation: Darriba D, Taboada GL, Doallo R and Posada D. 2012.
          "jModelTest 2: more models, new heuristics and parallel computing".
          Nature Methods 9(8), 772.
--------------------------------------------------------------------------------

Reading data file "Actin_1_MAFFT.fas"... failed.

Reading data file "ITS_1.fas"... OK.
  number of sequences: 1491
  number of sites: 532


---------------------------------------------------------------
*                                                             *
*        COMPUTATION OF LIKELIHOOD SCORES WITH PHYML          *
*                                                             *
---------------------------------------------------------------

::Settings::
```

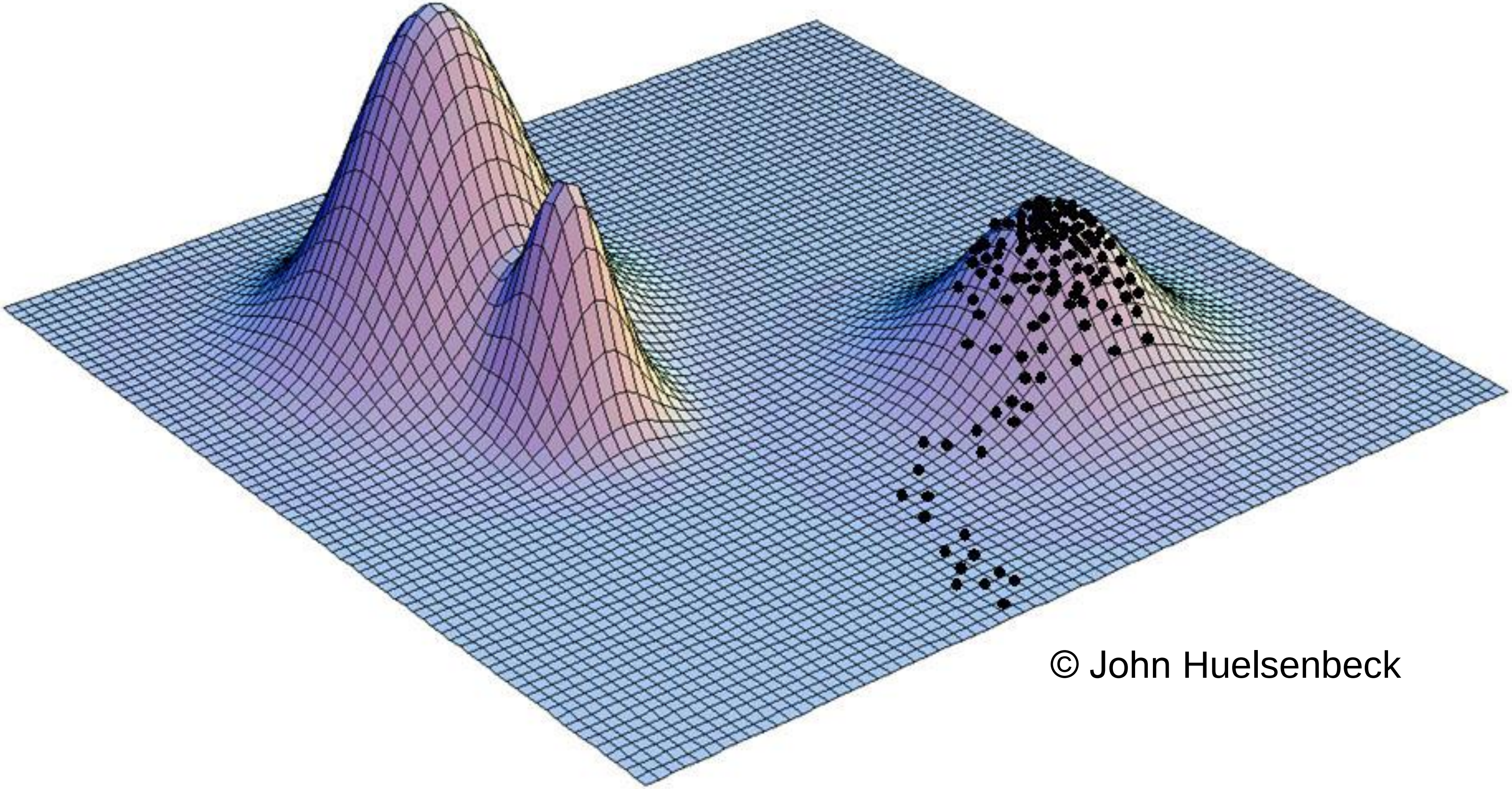
Likelihood scores not available ITS_1.fas

Best partitioning

- Partition Finder (Lanfear et al.) – selecting best-fit partitioning schemes and models of evolution
  - for all partitions simultaneously
  - merge partitions with same model into one
  - requires Python 2.7
  - alignment in phylip format
  - configuration file

```
# ALIGNMENT FILE #
alignment = test.phy;

# BRANCHLENGTHS #
branchlengths = linked;

# MODELS OF EVOLUTION #
models = all;
model_selection = aicc;

# DATA BLOCKS #
[data_blocks]
Gene1_pos1 = 1-789\3;
Gene1_pos2 = 2-789\3;
Gene1_pos3 = 3-789\3;

# SCHEMES #
[schemes]
search = greedy;
```

Bayesian inference - MrBayes



© John Huelsenbeck

Bayesian inference - MrBayes

```
begin mrbayes;



charset rbcL1       = 3-1179 \ 3;
charset rbcL2       = 1-1180 \ 3;
charset rbcL3       = 2-1181 \ 3;
charset ITS12       = 1182-1525 1683-1902;
charset RNA         = 1526-1682;

partition marker = 5:rbcL1,rbcL2,rbcL3,ITS12,RNA;

set partition = marker;

lset applyto=(1)   nst=6  rates=gamma;

lset applyto=(2,5) nst=1  rates=equal;

lset applyto=(3,4) nst=2  rates=gamma;

prset applyto=(1,3,4)  statefreqpr=dirichlet(1,1,1,1);

prset applyto=(2,5)  statefreqpr=fixed(equal);

prset applyto=(all) ratepr=variable;

unlink statefreq=(all) revmat=(all) tratio=(all) shape=(all)
pinvar=(all);

mcmc ngen=5000000 samplefreq=100 nchains=4;
end;
```

```
begin mrbayes;

charset ITS1             = 1-152;
charset ITS2             = 319-519;
charset RNA              = 153-318;
charset intron1          = 520-727;
charset exon             = 728-850;
charset intron2          = 851-1142;

partition vse = 6:ITS1,ITS2,RNA,intron1,exon,intron2;

set partition = vse;

lset applyto=(all)   nst=mixed  rates=gamma;

prset applyto=(all)  statefreqpr=dirichlet(1,1,1,1);

prset applyto=(all) ratepr=variable;

unlink statefreq=(all) revmat=(all) tratio=(all) shape=(all) pinvar=(all);

mcmc ngen=5000000 samplefreq=100 nchains=4;
end;
```

Parsimony - PAUP

```
[maximum parsimony block]
begin paup;
log start=yes file=MP.log replace=yes;
set autoclose=yes warnreset=no increase=auto;
set criterion=parsimony;
hsearch;
savetrees brlens=yes file=treeMP.tre replace=yes;
contree /strict=yes majrule=yes treefile=contree.tre replace=yes;
log stop;
end;
```

```
[bootstrap MP block]
log start=yes file=MPboot.log replace=yes;
bootstrap search=heuristic nreps=100 conlevel=50;
savetrees from=1 to=1 file=MPboot.tre savebootp=nodelabels maxdecimals=1 replace=yes;
log stop;
end;
```

ML - RAxML

Basic command line parameters

-m	substitution model
-p	random seed
-t	starting tree (if not specified, parsimony tree is generated using randomized stepwise addition)
-s	input file (phylip or fasta)
-#	number of replicates
-n	suffix for resulting files

Bootstrapping

a) find best ML tree (best-scoring tree)

```
raxmlHPC -m GTRGAMMA -p 12345 -# 20 -s dna.phy -n bestML
```

- generate 20 trees, best saved to RAxML_bestTree.bestML

b) compute bootstrap replicates

```
raxmlHPC -m GTRGAMMA -p 12345 -b 12345 -# 100 -s dna.phy -n boot
```

- generate 100 bootstrap matrices
- trees generated to RAxML_bootstrap.boot

c) bootstrapu values mapped onto best ML tree

```
raxmlHPC -m GTRGAMMA -p 12345 -f b -t RAxML_bestTree.bestML -z RAxML_bootstrap.boot -n finalboot
```

- two files generated: RAxML_bipartitions.finalboot (bootstrap values as nodes) and RAxML_bipartitionsBranchLabels.finalboot (bootstrap values above branches)

Rapid bootstrap

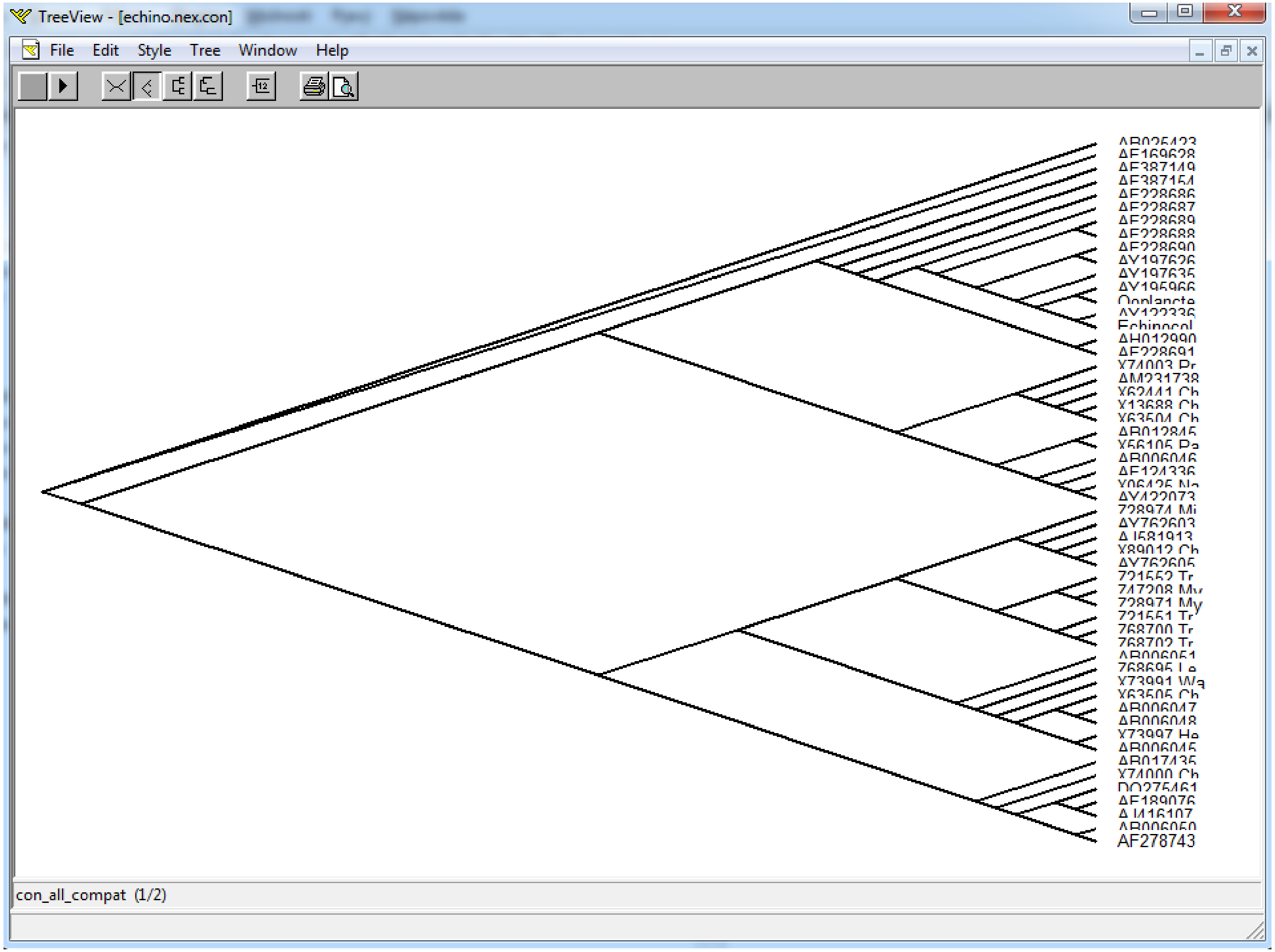
- much faster than standard bootstrap
- complete analysis (ML search+ bootstrapping) in one step

```
raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -# 100 -s dna.phy -n rbs
```

- RAxML_bipartitions.rbs – best ML tree with mapped bootstrap replicates

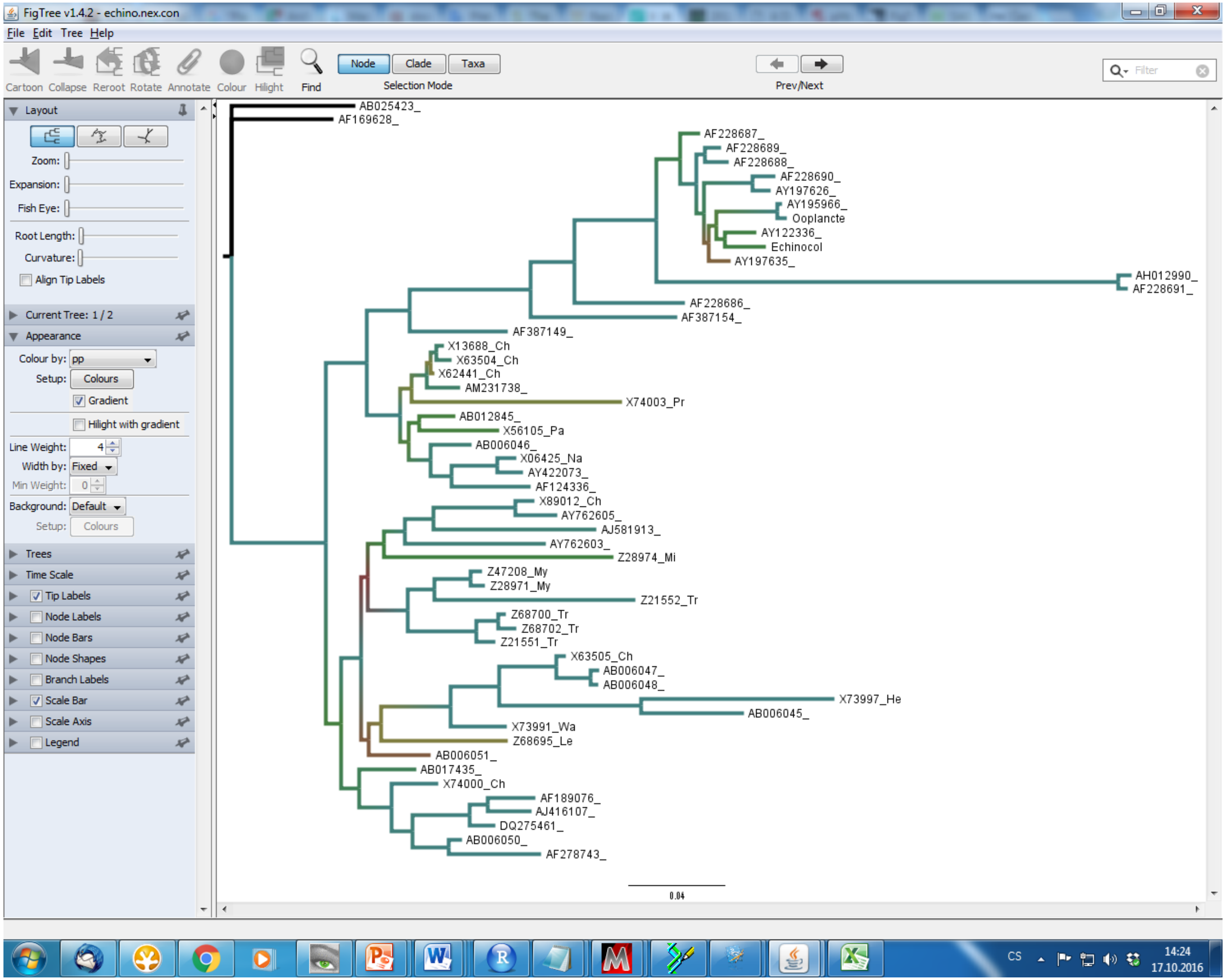
Tree Manipulation

- TreeView - http://en.bio-soft.net/tree/TreeView.html/



Tree Manipulation

- FigTree - http://tree.bio.ed.ac.uk/software/figtree/



FigTree v1.4.2 - echino.nex.con

Current Tree: 1 / 2

Colour by: pp

☑ Gradient

☐ Hilight with gradient

Line Weight: 4

Width by: Fixed

Min Weight: 0

Background: Default

☑ Tip Labels

☐ Node Labels

☐ Node Bars

☐ Node Shapes

☐ Branch Labels

☑ Scale Bar

☐ Scale Axis

☐ Legend

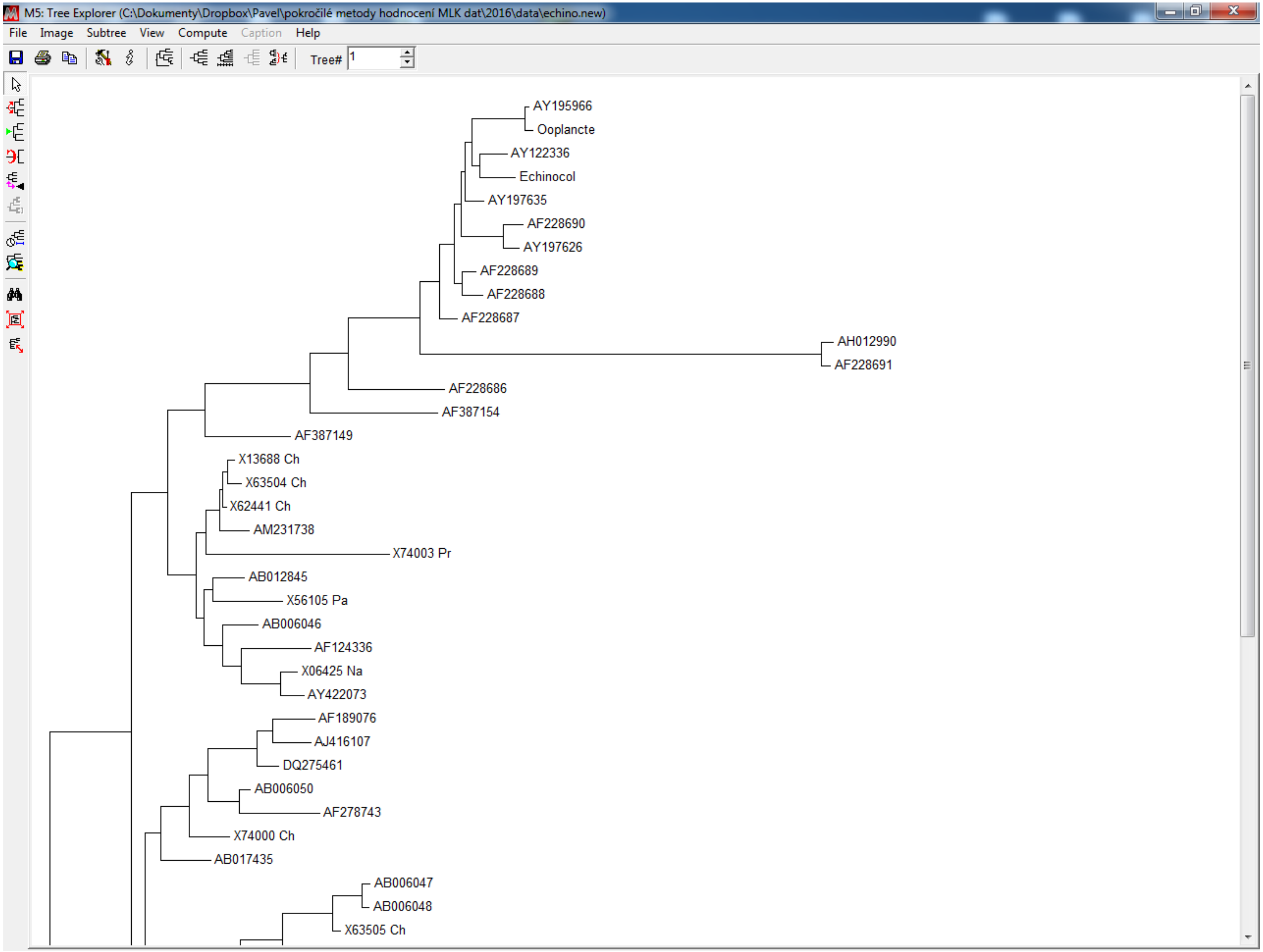
☐ Align Tip Labels

FigTree

Tree Manipulation

- MEGA - http://www.megasoftware.net/

MEGA Molecular Evolutionary Genetics Analysis



M5: Tree Explorer (C:\Dokumenty\Dropbox\Pavel\pokročilé metody hodnocení MLK dat\2016\data\echino.new)

Excercises:

Compare phylogenetic trees using:

- alignments with different amount of missing data
- alignments obtained by different MAFFT options
- original and minimized alignments after cleaning poorly aligned regions
- indel coding
- simple and complex evolutionary models
- partitioned and un-partitioned concatenated datasets
- different inferences (BI, ML, parsimony)