

# Tips for Phylogenomics

Ehsan Kayal  
Postdoctoral Fellow



*Phylogenetic trees are everywhere!*

Molecular Phylogenetic and Phylogenomic Approaches in Studies of Lichen Systematics and Evolution.

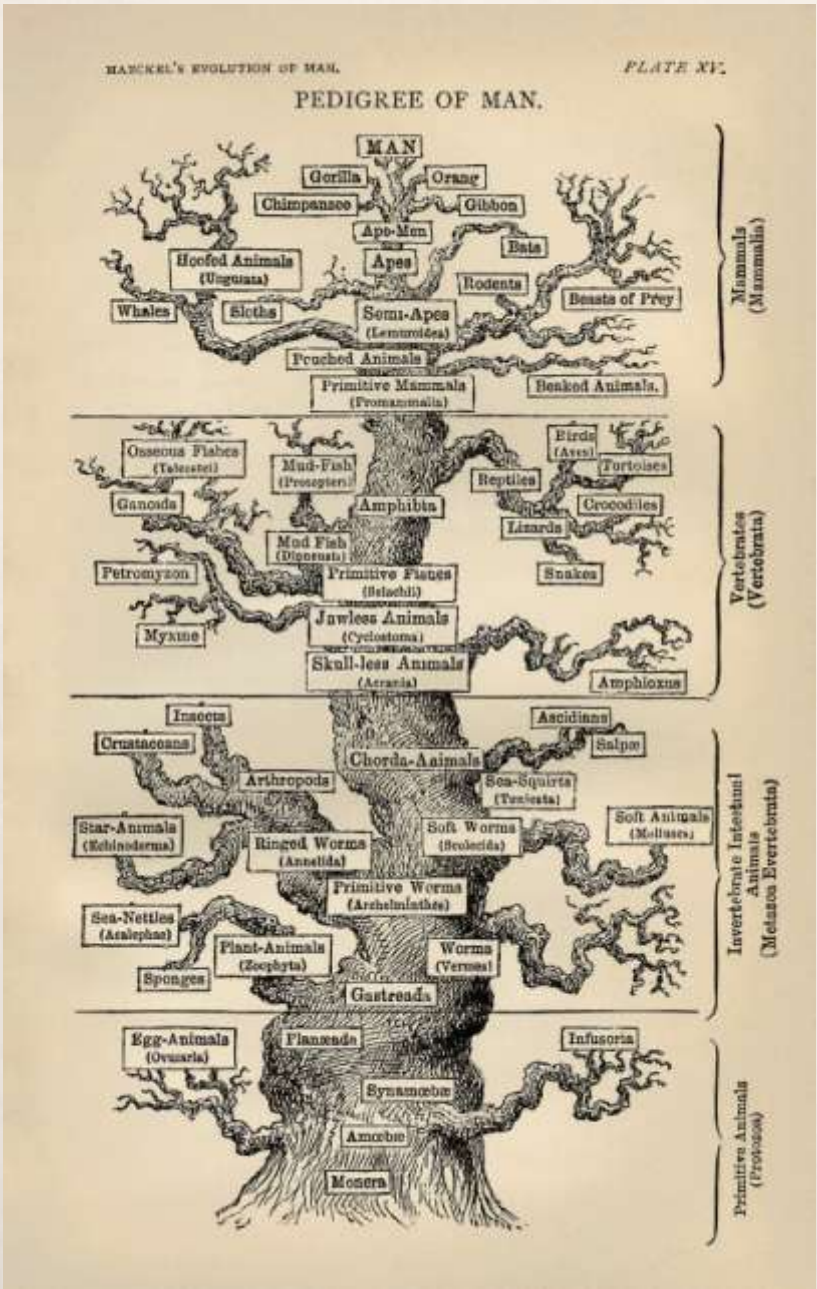
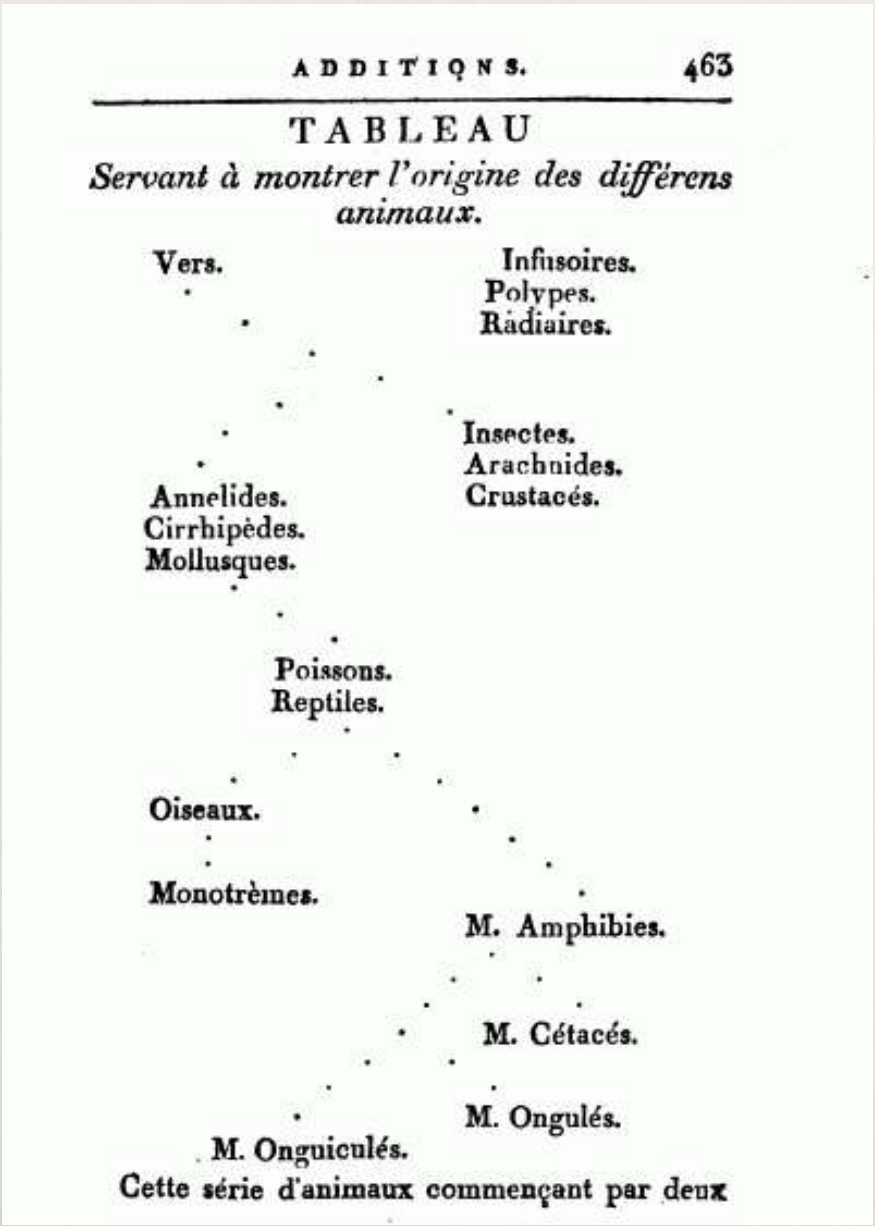
Divakar & Crespo, Recent Advances in Lichenology 2015: 45-60.

Phylogenetic tree shapes resolve disease transmission patterns.

Colijn & Gardy, EMPH 2014 (1): 96-108.

The use of phylogenetic analysis as evidence in criminal investigation of HIV transmission.

Bernard et al., HIV Forensics 2007.





# Four major frameworks

① Parsimony	➔	PAUP*
② Distance matrix	➔	PHYLIP
③ Maximum Likelihood	➔	RAxML
④ Bayesian	➔	MrBayes

*[evolution.genetics.washington.edu/phylip/software.html](http://evolution.genetics.washington.edu/phylip/software.html)*

# Steps to getting a phylogenetic tree

1. Get data
2. Isolate homologous genes
3. Build alignment
4. Infer a model of sequence evolution
5. Run phylogenetic analyses
6. Interpret results

# Data collection and filtering

- Before sequencing, check for available data (NCBI, DDBJ, EMBL, project-specific repositories)
- Gene selection (OrthoDB, OrthoMCL, ExPASy, BLASTO, etc...)  
[omictools.com/orthologous-groups-c419-p1.html](http://omictools.com/orthologous-groups-c419-p1.html)
- Check sequences (see Alignment)

# Sequence alignment

List of alignment software: [en.wikipedia.org/wiki/List\\_of\\_sequence\\_alignment\\_software](https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

## MAFFT v.7 ([mafft.cbrc.jp/alignment/server/](http://mafft.cbrc.jp/alignment/server/))

### MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences



#### Download version

[Mac OS X](#)

[Windows](#)

[Linux](#)

[Source](#)

[Online version](#)

[Alignment](#)

[mafft --edit Updated!](#)

[Merge Updated!](#)

[Phylogeny](#)

[Rough tree](#)

[Merits / limitations](#)

[Algorithms](#)

[Tips](#)

[Benchmarks](#)

[Feedback](#)

There are a few cases incompatible with **Avira Antivirus**. If this service does not work while running Avira, please temporarily disable Avira. Both protein and DNA data can be affected by this problem. (2015/Apr/14)

All jobs are reset at 4:00AM (JST) every Sunday.

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:

Paste protein or DNA sequences in fasta format. [Example](#)

A large, empty rectangular text input area with a light gray border, intended for pasting protein or DNA sequences in FASTA format.

or upload a plain text file:  No file selected.

☐ Use structural alignment(s)

☐ Allow unusual symbols (Selenocysteine \*U\*, Inosine \*I\*, non-alphabetical characters, etc.) [Help](#)

UPPERCASE / lowercase:

☐ Same as input

☒ Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences:

☒ Same as input

☐ Adjust direction according to the first sequence (accurate enough for most cases) **Beta**

☐ Adjust direction according to the first sequence (only for highly divergent data; extremely slow) **Beta**

Output order:

## Multiple sequence alignment

### Input:

Paste protein or DNA sequence

or upload a **plain text** file:

- ☐ Use structural alignment
- ☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

### UPPERCASE / lowercase:

- ☐ Same as input
- ☒ Amino acid → UPPERCASE

### Direction of nucleotide sequence:

- ☒ Same as input
- ☐ Adjust direction according to reference
- ☐ Adjust direction according to template

### Output order:

- ☐ Same as input
- ☒ Aligned

### Notify when finished (optional)

Email address:

- ☒ Use structural alignment(s)

### Structural alignment 1 (optional):

Paste an alignment in fasta format. [Example](#)

These sequences will be aligned with the 'input' sequences above, being used as a constraint.

### Structural alignment 2 (optional):

### Structural alignment 3 (optional):

### Structural alignment 4 (optional):

- ☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)



## 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 highly divergent ncRNAs with <200 sequences × <1,000 nucleotides) [Help](#)

## Parameters:

Scoring matrix for amino acid sequences:

Scoring matrix for nucleotide sequences:

[Switch it to '1PAM / κ=2' when aligning closely related DNA sequences.](#)

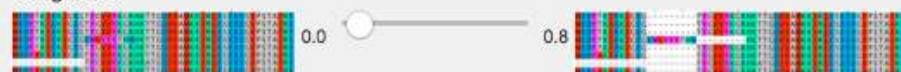
Gap opening penalty:  (1.0 - 3.0)

Offset value:  (0.0 - 1.0)

## 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 'Leave gappy regions'.

## Unalignlevel:



- ☒ Try to align gappy regions anyway  
☐ Leave gappy regions

Mafft-homologs (Collects homologs from SwissProt by BLAST and performs profile-based alignments; Protein only): [Help](#)

- ☐ On  
☐ Show homologs (if any)  
Number of homologs:  (5 - 200)  
Threshold: E =  (1e-5 - 1e-40)

Plot [LAST](#) hits (DNA only):

- ☒ The top sequence vs the others ☐ The longest sequence vs the others  
☒ Plot and alignment ☐ Plot only ☐ Alignment only

Threshold:

# MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

[Download version](#)

[Mac OS X](#)

[Windows](#)

[Linux](#)

[Source](#)

[Online version](#)

[Alignment](#)

[mafft --add](#) **Updated!**

[Merge](#) **Updated!**

[Phylogeny](#)

[Rough tree](#)

[Merits / limitations](#)

[Algorithms](#)

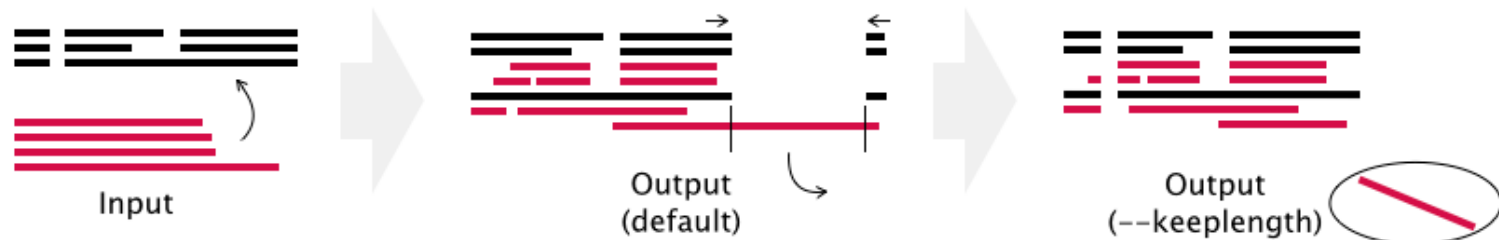
[Tips](#)

[Benchmarks](#)

[Feedback](#)

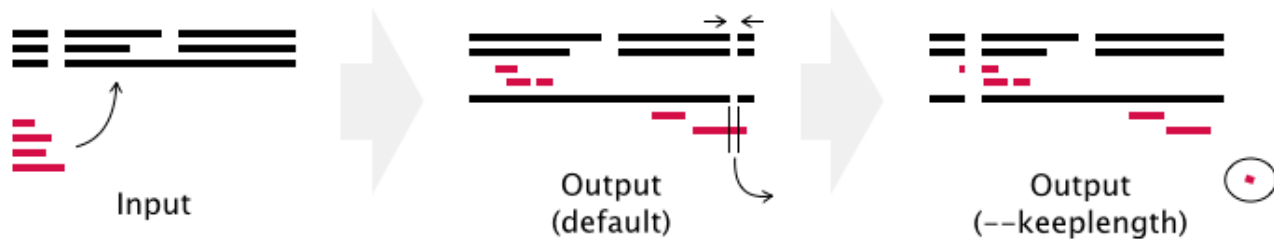
## **--add**

[Align full length sequences to an MSA](#)



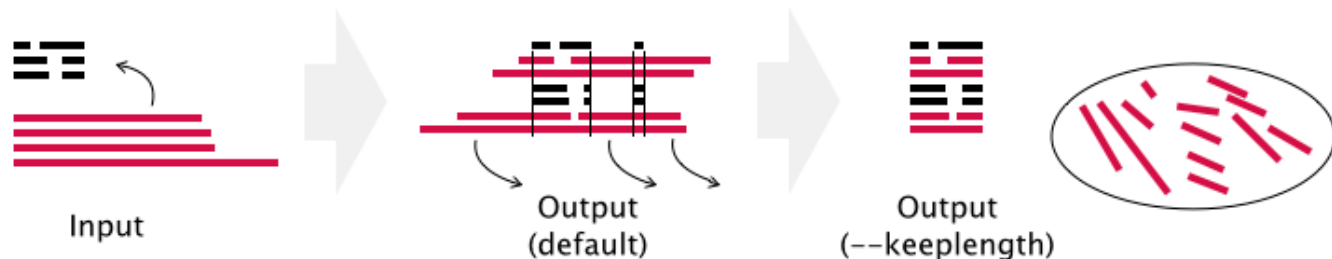
## **--addfragments**

[Align fragment sequences to an MSA](#)



## **--addlong** (experimental)

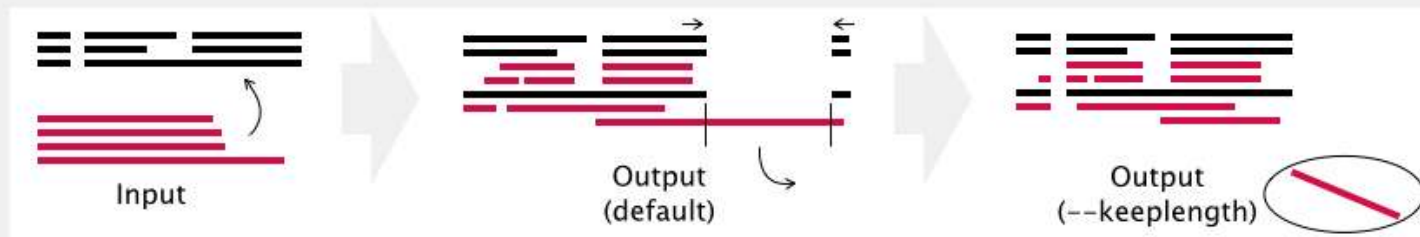
[Align long sequences to a short MSA](#)



Keep alignment length: **in Alpha Testing** (2015/May)

☐ Yes

With this option, insertions in the **new sequences** are deleted, to keep the alignment length the same as the **input alignment**.



A [correspondence table](#) between the positions in each **new sequence** and the positions in the alignment will also be returned.

Existing alignment: [Example](#)

Gaps (-) will be preserved.

or upload a plain text file:  No file selected.

[Clear](#)

**New sequence(s)** to be added to the above alignment: [Example](#)

Gaps (if any) will be removed.

or upload a plain text file:  No file selected.

[Clear](#)

☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

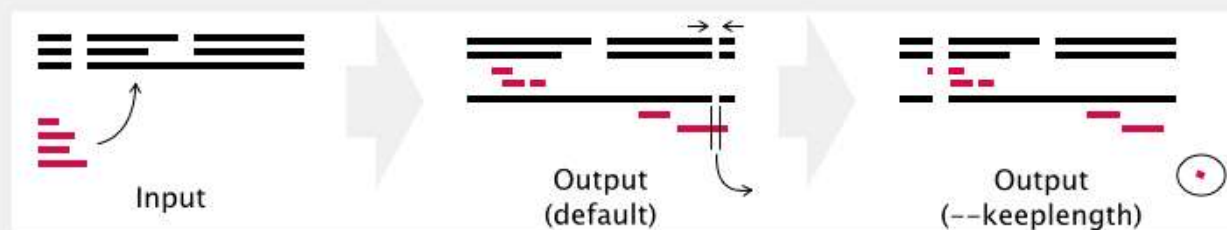
### Strategy:

- ☒ Auto (FFT-NS-1 or L-INS-i; depends on data size)
- ☐ FFT-NS-1 (Fast)
- ☐ G-INS-1 (Slow; uses all-pair global alignment)
- ☐ L-INS-1 (Slow; uses all-pair local alignment)

Keep alignment length: **in Alpha testing** (2015/May)

☐ Yes

With this option, insertions at the **fragmentary sequences** are deleted, to keep the alignment length the same as the **input alignment**.



A [correspondence table](#) between the positions in each **fragmentary sequence** and the positions in the alignment will also be returned.

Existing alignment: [Example](#)

Gaps (-) will be preserved.

or upload a plain text file:  No file selected.

[Clear](#)

**Fragmentary sequence(s)** to be added to the above alignment: [Example](#)

Gaps (if any) will be removed.

or upload a plain text file:  No file selected.

[Clear](#)

☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

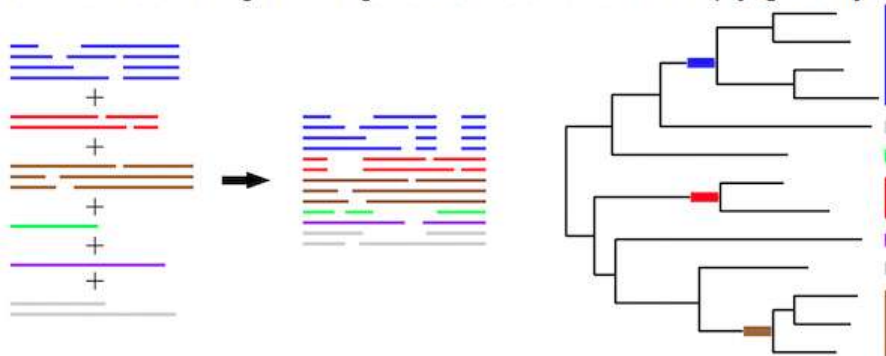
### Strategy:

- ☒ Auto (--multipair or --6merpair; depends on data size)
- ☐ --6merpair (Fast)
- ☐ --multipair --weighti 0 (Intermediate)
- ☐ --multipair (Accurate)

[Download version](#)[Mac OS X](#)[Windows](#)[Linux](#)[Source](#)[Online version](#)[Alignment](#)[mafft --add](#) **Updated!**[Merge](#) **Updated!**[Phylogeny](#)[Rough tree](#)[Merits / limitations](#)[Algorithms](#)[Tips](#)[Benchmarks](#)[Feedback](#)

## Merge two or more sub MSAs into a single MSA *In alpha testing* (2015/Jun) [Help](#)

Two or more sub MSAs are merged into a single MSA. Sub MSAs are assumed to be **phylogenetically separated** from each other. If it cannot be assumed, try [--add](#) or [--addfragments](#).



### Sub MSA a: [Example](#)

Gaps (-) will be preserved, except for gap-only sites.

or upload a plain text file:  No file selected. [Clear](#)

### Sub MSA b: [Example](#)

Gaps (-) will be preserved, except for gap-only sites.

or upload a plain text file:  No file selected. [Clear](#)

### 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)  
☐ L-INS-i (Very slow; recommended for <200 sequences with one conserved domain)  
☐ 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 RNA)

### Parameters:

Scoring matrix for amino acid sequences:  [v](#)

Scoring matrix for nucleotide sequences:  [v](#)

† Switch it to '1PAM / κ=2' when aligning closely related DNA sequences.

Gap opening penalty:  (1.0 - 3.0)

Offset value:  (0.0 - 1.0)



# Protein genes have to be kept in frame!

	510	520	530	540		550	560	570	580	590	600
Consensus	TTTARGWGCWTKTGATATTWCWTC AATTATT RAY --- GAAGATTTNAAATTC TATTGGTAAAAATTWTAATTATATTA --- TCTTTWTT										
Identity	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>										
1. Hydractinia ...	TATTAAGGAATTAATGATTATTCAAATT TTTGAT --- AATTATGTACAAAAAATAGGAATAATTATATATATTTTA --- TCACTTTT										
2. JN642330 ...	AACGGGTACCCCTCTGAATCTCTTCATCTACTGACTGTTTAGCAAACACTTACTCTCTCTGTGACCTTATAGCTTCCCTTATTTGTTCGGT TTT										
3. JN700934 ...	TTTAGGAGTTTGTGATATACAGTAATTATTAAT --- GAAGATT --- ATTCCTATGGGAAAAATTTATTAATATGATA --- GCCTTATT										
4. JN700944 ...	CTCGAGGGGCATCTAAGCTTTCATGGGGGGTTAAACCAATTGATTAGTTGGGGAATCTTTTCCCTGTACTGTTAATTACAGGG --- ACTTTATT										
5. JN700947 (...)	TTTAAGAGATTATATTTAA CAGTAACATATGGAT --- AAAAAATTTAATT CAGATTCTCTCTTTTGTGATTGTTTA --- TCTTTATT										
6. JN700949 ...	CACAGGCTCTTGTGATATAGCCTCAC TGTCCGGAC --- GGGTATT --- TGTCTTTTGAAAAGGTAAATGGTAATAGCA --- GCTTTGTT										
7. Mitocomella...	TTTA --- TCTGACGTGTTATATAGTCTAAT --- GATGTTTTGATTCTATTGGGTCAATATTAATTTATTG --- TCTTTTGC										
8. NC_00844...	TTCAAGGTAGCTGTGAATTAATCTACTTTGATAACC --- CAGGAAT --- ATACCCCTGAGAAAGTTCTTATTTTAGTA --- GCATTATT										
9. NC_01021...	TCTAAATTTATATAATTTTATTCTAAATAATTCC --- GTTTCATTTTACTCTTTAGTTATATTATTTTAATATTA --- TCAC TTT										

	510	520	530	540	550	560	570	580	590	600														
Consensus	TTTARGWGCWTKTGATATTWCWTC	CAATTATT	RAY	---	GAAGATT	TNAAAT	TCTATT	GGTAAA	ATTWTA	ATTATATTA	---	TCTTTWTT												
Frame 1	L X A C/F D I T/S S I I N/D	---	---	---	E D L/F N S I G K I I/L I I L	---	---	---	---	---	---	S L/F F												
Identity																								
1. Hydractinia ...	TATTA	AAAGAA	TTAATG	ATT	TATTC	AAATTT	TGAT	---	---	---	---	TCAC	TTT											
Frame 1	I K E L M I Y S N F D	---	---	---	N Y V T K I G I I I I L	---	---	---	---	---	---	S L F												
2. JN642330 ...	AACGGG	TACCC	CTCTGA	ATCTCT	TCATCT	ACTGACT	GTTTAG	CAAA	CACCTT	ACTCTT	CTCTGT	GACCTT	ATAGCT	TCCTTA	TTTGT	TCCG	TTT							
Frame 1	T G T L W I S S S T D C L A N T Y S S L W P Y S F L I C	---	---	---	A N T Y S S L W P Y S F L I C	---	---	---	---	---	---	---	---	---	---	S V F								
3. JN700934 ...	TTTAGG	AGTTT	GTGAT	TATAC	AGTA	AAATTA	TAAAT	---	---	---	---	GCCT	TATT	---	---	---	---							
Frame 1	L G V C D Y T V I I N	---	---	---	E D --- Y S M G K L L I M I	---	---	---	---	---	---	A L L	---	---	---	---	---							
4. JN700944 ...	CTCGAG	GGGCAT	CTAAGC	TTTCAT	GGGGGG	TTAAAC	CATTGA	TTAGT	TGGGG	AATCTT	TTCCTG	TACTGT	TAATTA	CAGGG	---	---	ACTTTATT							
Frame 1	T R G H L S F H G G L N H W L V G E S F P V L L I T G	---	---	---	L V G E S F P V L L I T G	---	---	---	---	---	---	---	---	---	---	---	T L F							
5. JN700947 (...)	TTTAAG	AGATT	TATAT	TTAAC	AGTA	AACTAT	GGAT	---	---	---	---	AAAA	AATTA	ATTCAG	ATT	TCTTCT	TTTTTT	TGATT	TGTTTA	---	---	TCTTTATT		
Frame 1	L R D L Y L T V T M D	---	---	---	K N L I Q I S S F L I C L	---	---	---	---	---	---	S L F	---	---	---	---	---	---	---	---	---	S L F		
6. JN700949 ...	CACAGG	CTCTT	GTGAT	ATAGC	TACCT	CACGT	CGGAC	---	---	---	---	GGGT	ATT	---	TGTC	TTT	TGAAA	AGGTAA	GGTA	AATAGCA	---	---	GCTTTGTT	
Frame 1	T G S C D I A S L S D	---	---	---	G Y --- L S F E K V M V I A	---	---	---	---	---	---	A L L	---	---	---	---	---	---	---	---	---	---	A L L	
7. Mitocomella...	TTTA	---	---	---	TCTGAC	CGTG	TATAT	AGTTCT	TAAT	---	---	GATG	TTTTG	ATTG	TATT	TGGGT	CAAT	ATT	TAATTT	TATTG	---	---	TCTTTTGC	
Frame 1	L --- S D V L Y S S N	---	---	---	D V L I V I G S I L I L L	---	---	---	---	---	---	S F A	---	---	---	---	---	---	---	---	---	---	S F A	
8. NC_00844...	TTCA	GGTAG	CTGTGA	AAATT	ACTAC	TTTGAT	AACC	---	---	---	---	CAGGA	AT	---	ATAC	CCCTG	GAGAA	GGTT	CTTAT	TTTAGTA	---	---	GCA	TATT
Frame 1	S G S C E I T T T L I T	---	---	---	Q E --- Y T P E K V L I L V	---	---	---	---	---	---	A L L	---	---	---	---	---	---	---	---	---	---	A L L	
9. NC_01021...	TCTA	AAATTA	ATAAT	TTTAT	TCT	AAATAA	TCC	---	---	---	---	GTTT	CA	TTT	ACTCT	TTT	AGTT	ATAT	ATTTA	ATATTA	---	---	TCAC	TTT
Frame 1	L N L Y N F I L N N S	---	---	---	V S F Y S F S Y I I L I L	---	---	---	---	---	---	S L F	---	---	---	---	---	---	---	---	---	---	S L F	

## INPUT FILE 1

Put your multiple sequence al

- The server automatical
- In the CLUSTAL form:
- [How to specify stop co](#)

or

upload

nad2 translation alignmr

## INPUT FILE 2

Put your DNA (or mRNA) se

or

upload

nad2.fasta

## Option setting

### 1. Codon table:

- ☐ Universal code (NCBI: transl\_table=1)
- ☐ Vertebrate mitochondrial code (NCBI: transl\_table=2)
- ☐ Yeast mitochondrial code (NCBI: transl\_table=3)
- ☒ Mold, Protozoan, and Coelenterate Mitochondrial code and Mycoplasma/Spiroplasma code (NCBI: transl\_table=4)
- ☐ Invertebrate mitochondrial code (NCBI: transl\_table=5)
- ☐ Ciliate, Dasycladacean and Hexamita nuclear code (NCBI: transl\_table=6)
- ☐ Echinoderm and Flatworm mitochondrial code (NCBI: transl\_table=9)
- ☐ Euplotid nuclear code (NCBI: transl\_table=10)
- ☐ Bacterial, archaeal and plant plastid code (NCBI: transl\_table=11)
- ☐ Alternative yeast nuclear code (NCBI: transl\_table=12)
- ☐ Ascidian mitochondrial code (NCBI: transl\_table=13)
- ☐ Alternative flatworm mitochondrial code (NCBI: transl\_table=14)
- ☐ Blepharisma nuclear code
- ☐ Chlorophycean mitochondrial code
- ☐ Trematode mitochondrial code
- ☐ Scenedesmus obliquus mitochondrial code
- ☐ Thraustochytrium mitochondrial code

### 2. Remove gaps, inframe stop codons:

- ☒ No
- ☐ Yes
  - Calculate  $d_S$  and  $d_N$ :
    - ☒ No
    - ☐ Yes (valid only if the input is a pair of sequences)

### 3. Remove mismatches: (mismatched codons between protein and DNA)

- ☒ No
- ☐ Yes

### 4. Use only selected positions ('#' under the input alignment):

- ☒ No
- ☐ Yes

### 5. Output format:

- ☐ CLUSTAL
- ☐ PAML
- ☒ FASTA
- ☐ Codon with Amino acid

# PAL2NAL conversion allows maintaining the reading frame

	520	530	540	550	560	570	580	590	600
Consensus	TWACTGMATTGATT			AATN-TNNNRATGATKTTTATTCT			WTTGGTWAAAATTTTWATTATATTATCTTTWTTTTT		
Frame 1	I */L L N/H W L			I/M-X X W C/F L F Y			/F W */L N F N/Y Y I I F I/F F		
Identity									
1. Hydractinia...	TATTAAGAATTGAATGATTTATTCAATTTTGATAATTATGTTACAAAA						ATAGGAATAAATTATTATAAATTTTATCAC		
Frame 1	I K E L M I Y S N F D N Y V T K						I G I I I I I L S L F F		
2. JN642330...	AATCTCTTCATCTACT			GACTGTTTAGCAAACACTTACTCTCTCTCTGTGACCTTATAGCTTCCCTTATTGTGTCGGTT					
Frame 1	I S S S T			D C L A N T Y S S L W P Y S F L I C S V F F					
3. JN700934...	TTATACAGTAAATTATT				AATGAAGATTATTCT		ATGGGAAAATTTATTAATAAATGATAGCCCTTATTATT		
Frame 1	Y T V I I				N E D Y S		M G K L L I M I A L L F		
4. JN700944...	CTTTCATGGGGGGTTA			AACCATGTAGTTGGGGAACTCT			TTTCCCTGTACTGTTAATTACAGGGACTTTATTCTT		
Frame 1	F H G G L			N H W L V G E S			F P V L L I T G T L F F		
5. JN700947...	TTTAAGAGATTATATTTAAACAGTAACATATGGGATAAAAAATTTAATTCAG						ATTTCTTCTTTTGGATTGTGTTTATCTTTATTTT		
Frame 1	L R D L Y L T V T M D K N L I Q						I S S F L I C L S L F F		
6. JN700949...	TATAGCCTCACGTGTCG				GACGGGTAATTGTCT		TTTGAAAAGGTAATGGTAATAGCAGCTTTGTTATT		
Frame 1	I A S L S				D G Y L S		F E K V M V I A A L L F		
7. Mitocome...	TTTATCTGACGTGTTA		TATAGTTCCTAATGATGTTTGTGTT				ATTGGGTCGAATATTAAATTTATTGTCTTTTGCTTT		
Frame 1	L S D V L		Y S S N D V L I V				I G S I L I L L S F A F		
8. NC_00844...	AATTAATAATTTGATA			ACCCAGGAATATACC			CCTGAGAAGGTTCTTATTTTAGTAGCATATTGTT		
Frame 1	I T T L I			T Q E Y T			P E K V L I L V A L L F		
9. NC_01021...	TCTAAATTTATATAATTTTATTCTAAATAAATCCGTTTCATTTTACTCT						TTTAGTTATATTATTTTAATATTATCCTTT		
Frame 1	L N L Y N F I L N N S V S F Y S						F S Y I I L I L S L F F		



Direct nt alignment (MAFFT-Auto)

	510	520	530	540	550	560	570	580	590	600
Consensus	TTTARGWGCWTKTGATATTWCWTC AATTATT RAY									
Identity	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>									
1. Hydractinia ...	TATTAAGGAATTAATGATTTATTCAAATTTTGAT				AAATTATGTACAAAAATAGGAATTAATTATATATAATTTTA					TCACCTTTT
2. JN642330 ...	AACGGGTACCCCTCTGAATCTCTTCATCTACTGACTGTTTAG				CAAACACCTTACTCTTCTCTGTGACCTTATAGCTTCCCTTA				TTTGTTC	CGTTT
3. JN700934 ...	TTTAGGAGTTTGTGATTATACAGTAATTATTAAT				GAAGATT--ATTCCTATGGGAAAAATATTAATATAAGATA					GCCTTATT
4. JN700944 ...	CTCGAGGGCATCTAAGCTTTCATGGGGGGTTAAACCATTTGA				TTAGTTGGGGAATCTTTTCCCTGTACTGTTAATTACAGGG					ACCTTATT
5. JN700947 (...)	TTTAAGAGATTATATTTAACAGTAACATATGGAT				AAAAAATTTAATTTCAGATTCTCTCTTTTGTGATTGTGTTTA					TCTTTATT
6. JN700949 ...	CACAGGCTCTGTGATATAGCCTCACCTGTCGGAC				GGGTATT--TGTCTTTTGAAAAAGGTAATGGTAATAGCA					GCTTTGTT
7. Mitocomella...	TTTA-----TCTGACGTGTTATATAGTTCCTAAT				GATGTTTGTGATTATTTGGGTCAATATTAATTATTATTG					TCTTTTGC
8. NC_00844...	TTCAAGGTAGCTGTGAATAATTACTTTGATAACC				CAGGAAT--ATACCCCTGAGAAAGTTCTTATTTTAGTA					GCAATTATT
9. NC_01021...	TCTAAATTTATATAATTTTATTCTAAATAATTCC				GTTTCAATTTACTCTTTAGTTATATTATTTTAAATATTA					TCACTTT

PAL2NAL converted alignment

	520	530	540	550	560	570	580	590	600	610
Consensus	TWTWACTGMATTGATT--AATN-TNNNRATGATKTTTATTCT									
Identity	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>									
1. Hydractinia...	TATTAAGGAATTAATGATTTATTC A A A A A A						ATAGGAATTAATTATATATAATTTTATCACTTTTTTTTAA			
2. JN642330_...	AATCTCTTCATCTACT--GACTGTTTAGCAAACACCTTACTCT						TCTCTGTGACCTTATAGCTTCCCTTATTTGTTCCGTTTTTTTCAA			
3. JN700934_...	TATACAGTAATTATTT--AATGAAGATTATTTCT						ATGGGAAAAATATTAAATAATGATAGCCCTTATTATTAA			
4. JN700944_...	CTTTCATGGGGGTTA--AACCATGATTAGTTGGGGAATCT						TTTCCCTGTACTGTTAATTACAGGGACTTTATTCTTTAA			
5. JN700947_...	TTTAAGAGATTATATTTAACAGTAACATATGGATAAAAAATTTAATTCAG						ATTTCCTCTTTTTTGTGTTTATCTTTATTTTTTTAA			
6. JN700949_...	TATAGCCTCACCTGTCG--GACGGGTATTTGTCT						TTTGAAAAAGGTAATGGTAATAGCAGCTTTGTTATTCAA			
7. Mitocomella...	TTTATCTGACGTGTTA--TATAGTTC A A T G A T G T T G A T T G T T						ATTGGGTCAATATTAAATTTATTTGTCTTTTGC			
8. NC_00844...	AATTACTACTTTGATA--ACCCAGGAATATACC						CCTGAGAAAGGTTCTTATTTTAGTATGCAATTATTGTTTAA			
9. NC_01021...	TCTAAATTTATATAATTTTATTCTAAATAATTCCGTTTCATTTTACTCT						TTTAGTTATATTATTTTAAATATTTATCACTTTTTTTCAA			

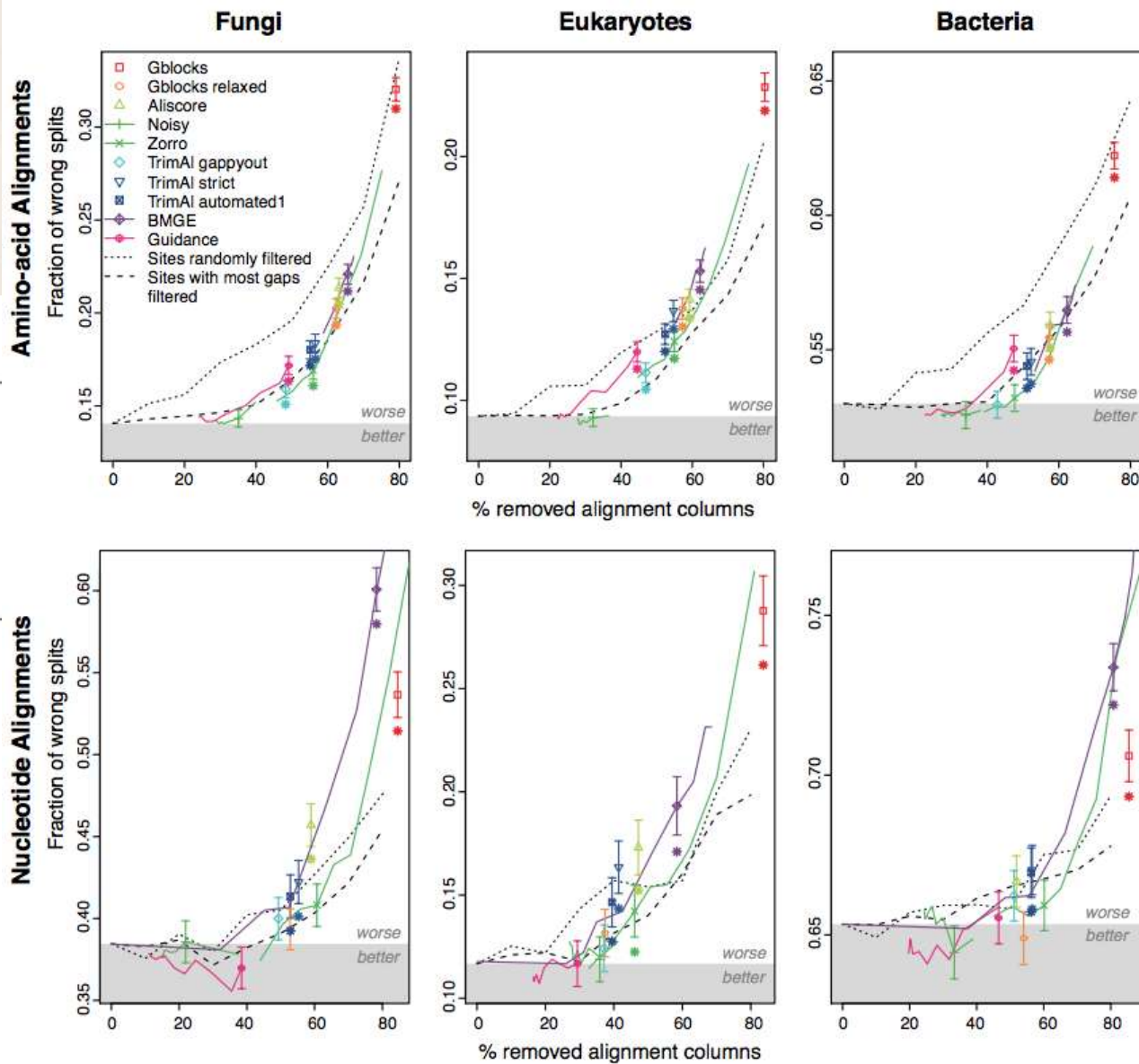


Figure 2: Alignment filtering generally yields poorer phylogenetic trees. Depicted here are results with the enriched species-tree discordance test on amino-acid (top) and nucleotide (bottom) alignments from three taxonomic ranges. The measure of error is the average Robinson-Foulds distance between the reference trees and trees reconstructed from Prank+F alignments filtered by the various approaches. Trees were reconstructed using PhyML. Filtered alignments improving over unfiltered alignment fall in the grey region.

ference

avera and Castresana  
(2007)

ella-Gutiérrez et al.  
(2009)

ss et al. (2008)

ck et al. (2010)

scuolo and Gribaldo  
(2010)  
et al. (2012)

n et al. (2010)



# Models of sequence evolution

## Published models:

✧ jmodeltest v.2 ([code.google.com/p/jmodeltest2/](http://code.google.com/p/jmodeltest2/))

DNA

- JC, F81, K80, HKY, TrN, TrM.... SYM, GTR
- +I, +G, +I+G and +F parameters.
- AIC, BIC, AICc, DT; BIONJ/ML

✧ prottest v.3 ([code.google.com/p/prottest3/](http://code.google.com/p/prottest3/))

amino acids

- WAG, Dayhoff, JTT, mtREV, MtMam, MtArt... Blosum62, LG
- +I, +G, +I+G and +F parameters.
- AIC, BIC, AICc, DT; BIONJ/ML

## Custom models:

➔ Many phylogenetic program allow custom-made models

# Some phylogenetic programs

[evolution.genetics.washington.edu/phylip/software.html](http://evolution.genetics.washington.edu/phylip/software.html)

[en.wikipedia.org/wiki/List\\_of\\_phylogenetics\\_software](http://en.wikipedia.org/wiki/List_of_phylogenetics_software)

- Parsimony:
  - PAUP\* ([paup.csit.fsu.edu/](http://paup.csit.fsu.edu/))
  - MEGA ([megasoftware.net/](http://megasoftware.net/))
  - TNT ([lillo.org.ar/phylogeny/tnt/](http://lillo.org.ar/phylogeny/tnt/))
- Maximum Likelihood:
  - RAxML ([sco.h-its.org/exelixis/web/software/raxml/index.html](http://sco.h-its.org/exelixis/web/software/raxml/index.html))
  - PAML ([abacus.gene.ucl.ac.uk/software/paml.html](http://abacus.gene.ucl.ac.uk/software/paml.html))
  - PhyML ([atgc-montpellier.fr/phyml/](http://atgc-montpellier.fr/phyml/))
  - GARLI ([bio.utexas.edu/faculty/antisense/garli/garli.html](http://bio.utexas.edu/faculty/antisense/garli/garli.html))
  - Treefinder ([treefinder.de/](http://treefinder.de/))
- Bayesian:
  - MrBayes ([mrbayes.sourceforge.net/](http://mrbayes.sourceforge.net/))
  - BEAST ([beast.bio.ed.ac.uk/](http://beast.bio.ed.ac.uk/))
  - PhyloBayes ([megasun.bch.umontreal.ca/People/lartillot/www/index.htm](http://megasun.bch.umontreal.ca/People/lartillot/www/index.htm))

# RAxML v.8+

[sco.h-its.org/exelixis/resource/download/NewManual.pdf](http://sco.h-its.org/exelixis/resource/download/NewManual.pdf)

Data format: PHYLIP (.phy) or FASTA (.fa/ .fasta); Newick

## Model list:

NT: GTRGAMMA, GTRCAT

AA: DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, MTMAM, LG, MTART, MTZOA, PMB, HIVB, HIVW, JTTDCMUT, FLU, DUMMY, DUMMY2, LG4M, LG4X, PROT\_FILE, GTR\_UNLINKED, GTR

## Quick and dirty:

```
raxmlHPC-SSE -m MODEL -f ae -#NUM -p 12345 -x 12345 -s IN.phy -n RUN.out
```

Ex:

```
raxmlHPC-SSE -m GTRGAMMAIF -f ae -#100 -p 12345 -x 12345 -s align.phy -n raxml_run
```

Ex:

```
raxmlHPC-SSE -m PROTCATGTR -f ae -#100 -p 12345 -x 12345 -s align.phy -n raxml_run
```

# RAxML v.8+

## Advanced:

### 1. ML tree search:

```
raxmlHPC-SSE -m MODEL -d -f d/o -#NUM -p 12345 -s IN.phy -n Tree.out
```

Ex:

```
raxmlHPC-SSE -m GTRCAT -d -f d -#100 -p 12345 -s align.phy -n raxml.bestree
```

### 2. Bootstrapping:

```
raxmlHPC-SSE -m MODEL -b 12345 -#NUM -s IN.phy -n Boot.out
```

Ex:

```
raxmlHPC-SSE -m GTRCAT -#100 -b 12345 -s align.phy -n raxml.boot
```

### 3. Merging tree and support values:

```
raxmlHPC-SSE -f b -t Tree.out -z Boot.out -n RUN.out
```

Ex:

```
raxmlHPC-SSE -f b -t raxml.bestree -z raxml.boot -n raxml.run
```

# MrBayes v.3.2

[mrbayes.sourceforge.net/mb3.2\\_manual.pdf](http://mrbayes.sourceforge.net/mb3.2_manual.pdf)

Data format: NEXUS (.nex/.nxs)

```
#NEXUS
begin taxa;
  dimensions ntax=53;
  taxlabels
  Physalia_ESTs_Angel
  Physalia_physalia_SRR871528
  Rhizophysa_DLSI230
  Hydractinia_symbiolongicarpus_SRR1174275_SRR1174698
  Hydractinia_polyclina_SRR923509
;
end;

begin characters;
  dimensions nchar=2134;
  format datatype=dna missing=? gap=- interleave=yes;
  matrix
  Physalia_ESTs_Angel      AGTAAAGGCGATTGAATTTATAG-CG-T-GAGTACTGTGAAGGAATACCTTTTGTATAAT
  Physalia_physalia_SRR871528      AGTAAAGGCGATTGAATTTATAG-CG-T-GAGTACTGTGAAGGAATACCTTTTGTATAAT
  Rhizophysa_DLSI230      AGTAACGGCGAGTGAACCTTATAG-GA-T-AAGTACTGTGAAGGAATACCTTTTGTATAAT
  Hydractinia_symbiolongicarpus_SRR1174275_SRR1174698      AGTATTGGCAAAAGAAGAACTTTTAA-TT-A-TAGTACTGTGAAGGAATACCTTTTGTATAAT
  Hydractinia_polyclina_SRR923509      AGTATTGGCAAAAGAAGAACTTTTAA-TT-A-TAGTACTGTGAAGGAATACCTTTTGTATAAT

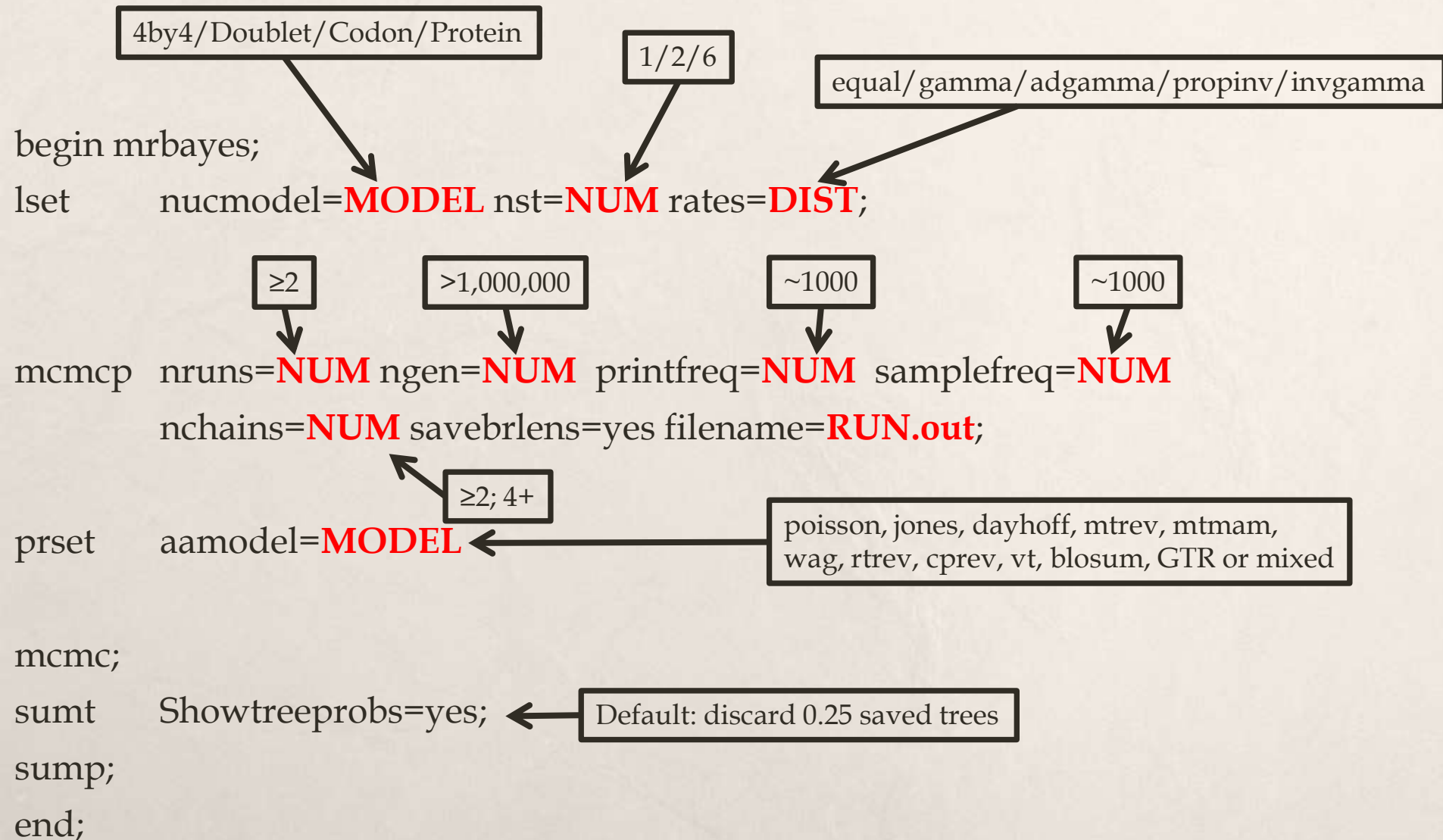
  Physalia_ESTs_Angel      CGTAACATAGTGGGGGG-----
  Physalia_physalia_SRR871528      CGTAACATAGTGGGGGGAAGGGAACCTTCTCCCTG
  Rhizophysa_DLSI230      CGTAACATAGTGGGGGGAAGGGAACCTTCTCCCTG
  Hydractinia_symbiolongicarpus_SRR1174275_SRR1174698      CGTAACATAGTGGGAGAAAGGGAACCTTTTCTCG
  Hydractinia_polyclina_SRR923509      CGTAACATAGTGGGRGAAAGGGAACCTTTTCTCG
;
end;

begin mrbayes;
  lset nucmodel=4by4 nst=6 ploidy=haploid rates=invgamma;
  mcmc nruns=2 ngen=500000 printfreq=1000 samplefreq=1000 nchains=4 savebrlens=yes filename=MrBayesGTRGIrRNAconcat54txGb;
  mcmc;
  sumt Showtreeprobs=yes;
  sump;
end;
```



# MrBayes v.3.2

MrBayes parameter block:



# MrBayes v.3.2

Ex:

- Nucleotides:

```
begin mrbayes;  
lset nucmodel=4by4 nst=6 ploidy=haploid rates=invgamma;  
mcmc nruns=2 ngen=5000000 printfreq=1000 samplefreq=1000 nchains=4 savebrlens=yes filename=mymbrun;  
mcmc;  
sumt Showtreeprobs=yes;  
sump;  
  
end;
```

- Amino acids:

```
begin mrbayes;  
lset nucmodel=Protein rates=invgamma;  
prset aamodel=fixed(gtr);  
mcmc nruns=3 ngen=10000000 printfreq=1000 samplefreq=1000 nchains=8 savebrlens=yes filename=mbaa;  
mcmc;  
sumt Showtreeprobs=yes;  
sump;  
  
End;
```

# PhyloBayes v.3.3f

[megasun.bch.umontreal.ca/People/lartillot/www/index.htm](http://megasun.bch.umontreal.ca/People/lartillot/www/index.htm)

Data format: PHYLIP(.phy)

- Starting a chain:

`pb -d align.phy -RATE -MODEL chainname` → At least 2 chains

Ex:      `pb -d align.phy -ratecat -cat -gtr pb-catgtr_1 &`  
          `pb -d align.phy -ratecat -cat -gtr pb-catgtr_2 &`

- Checking convergence:

`bpcomp -x BURN-IN NUM CHAIN1 CHAIN2 &`

Ex:      `bpcomp -x 100 10 pb-catgtr_1 pb-catgtr_2 &`

→ `bpcomp.bpdiff`: largest (**maxdiff**) and mean (**meandiff**) discrepancy

**maxdiff**<0.1 → good run

**maxdiff**<0.3 → acceptable run

→ `bpcomp.con.tre` = majority-rule posterior consensus tree

# PhyloBayes v.3.3f

## ○ Auto-stop run:

pb -d align.phy **-RATE -MODEL -nchain NUM CYCL MIN\_SIZ chainname**

Ex: pb -d align.phy -nchain 2 100 0.1 100 pb-cat

## ❖ Models:

-poi (F81, default); -jtt; -wag; -mtrev, -mtzoa, -mtart; -lg; -gtr; ~~-gtrm~~ or -rr (fixed)

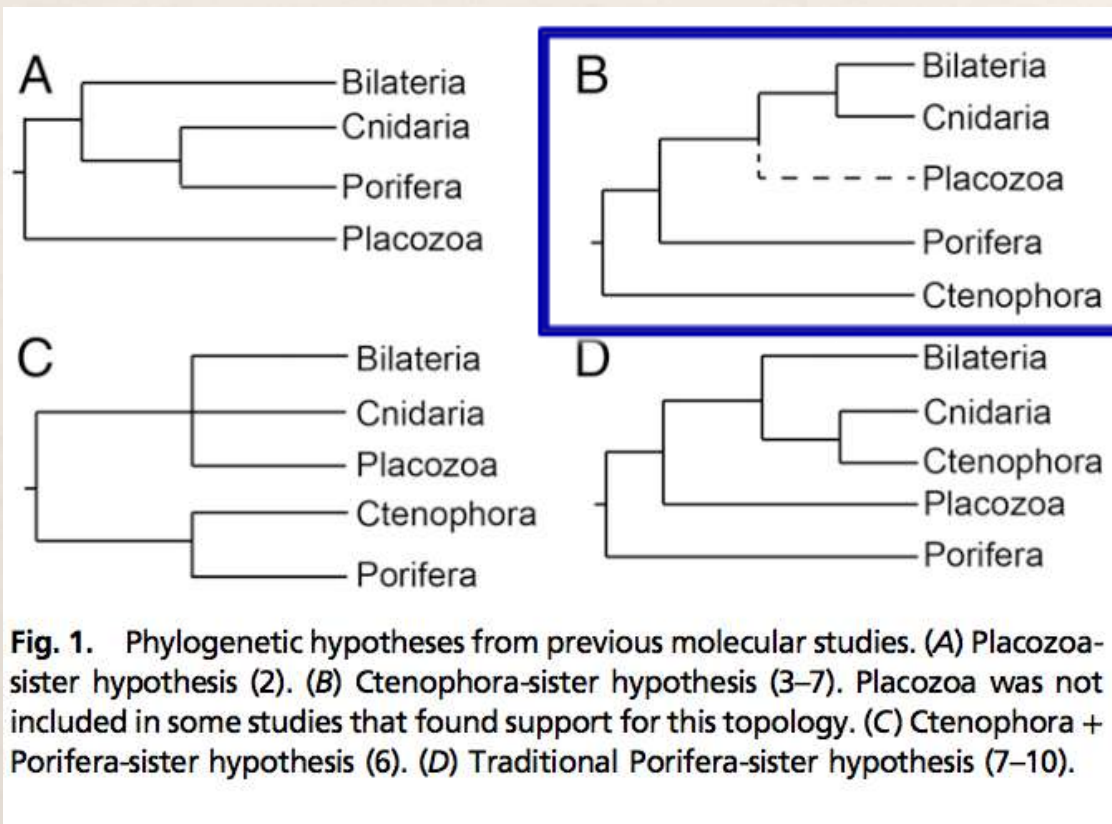


## ❖ Rates:

-ratecat; -uni; -cgam; -dgam (default)



# Interpreting phylogenetic results



Whelan *et al.* (2015) PNAS 112(18): 5773–5778

Past bias due to ribosomal proteins

➔ Data selection biased by automated orthologue finders?



Phylogenomics is a

OPEN ACCESS Freely available online

## Perspective

# Resolving Difficult Sequences Are Not

**Hervé Philippe<sup>1\*</sup>, Henner Brinkmann<sup>2</sup>, Gert Wörheide<sup>5,6</sup>, Denis Baurain<sup>3</sup>**

**1** Département de Biochimie, Centre Robert-Cedergren, Iowa State University, Ames, Iowa, United States of America, **2** Institut für Systematik, Zoologie, und Evolutionsbiologie, Universität München, München, Germany, **3** GeoBio-Center, Institute of Veterinary Medicine, University of Liège, Liège, Belgium, **4** Institut für Systematik, Zoologie, und Evolutionsbiologie, Universität München, München, Germany, **5** Institut für Systematik, Zoologie, und Evolutionsbiologie, Universität München, München, Germany, **6** Institut für Systematik, Zoologie, und Evolutionsbiologie, Universität München, München, Germany

In the quest to reconstruct the Tree of Life, researchers have increasingly turned to phylogenomics, the inference of phylogenetic relationships using genome-scale data (Box 1). Mesmerized by the sustained increase in sequencing throughput, many phylogeneticists entertained the hope that the incongruence frequently observed in studies using single or a few genes [1] would come to an end with the generation of large multigene datasets. Yet, as so often happens, reality has turned out to be far more complex, as three recent large-scale analyses, one published in *PLoS Biology*

its impact. Since taxon and gene sampling is being rapidly improved by the relentless progress in sequencing technology (even if obtaining well preserved and correctly identified specimens remains the limiting factor for several key taxa), full achievement of the ultimate goal of phylogenomics—i.e., accurate resolution of the Tree of Life—will primarily hinge on better procedures for the selection of orthologous and least saturated genes as well as on improved models of sequence evolution. In summary, while we certainly encourage the inclusion of neglected groups of organisms in large-scale sequencing studies (e.g., [2,3,46,48]), we consider at least as important that phylogeneticists engage in theoretical and bioinformatics developments that keep pace with sequencing technology to overcome these serious bottlenecks. This is essential to ensure that lessons learned from classical and molecular systematics are not forgotten in the phylogenomic era.

SYSTEMATIC BIOLOGY

more

Manuel<sup>4</sup>,

Systematic Biology,  
Paris 6, UMR  
Museum für  
nd Faculty of

organisms over  
should be  
al source of  
heless, even  
were not an  
sequences  
on does not  
the size of  
out so too is  
vs that non-  
e dominant  
statistically  
trees [12].



Questions?