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.

Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck

TABLEAU

Servant à montrer l'origine des différens animaux.

Vers.

Infusoires.
Polypes.
Radiaires.

Insectes.
Arachnides.
Crustacés.

Annelides.
Cirrhipèdes.
Mollusques.

Poissons.
Reptiles.

Oiseaux.

Monotrèmes.

M. Amphibies.

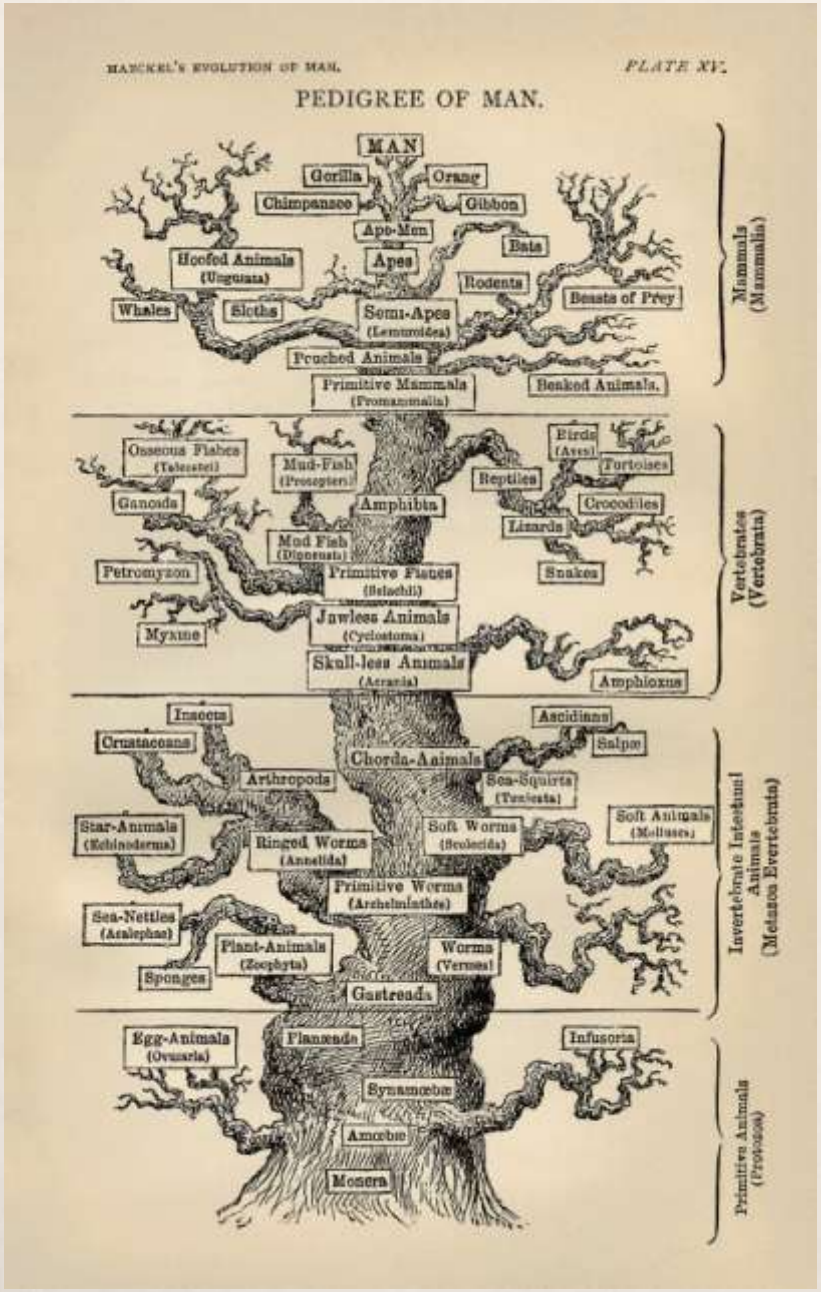
M. Cétacés.

M. Ongulés.

M. Onguiculés.

Cette série d'animaux commençant par deux

Ernst Heinrich Philipp August Haeckel



Four major frameworks

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

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…)

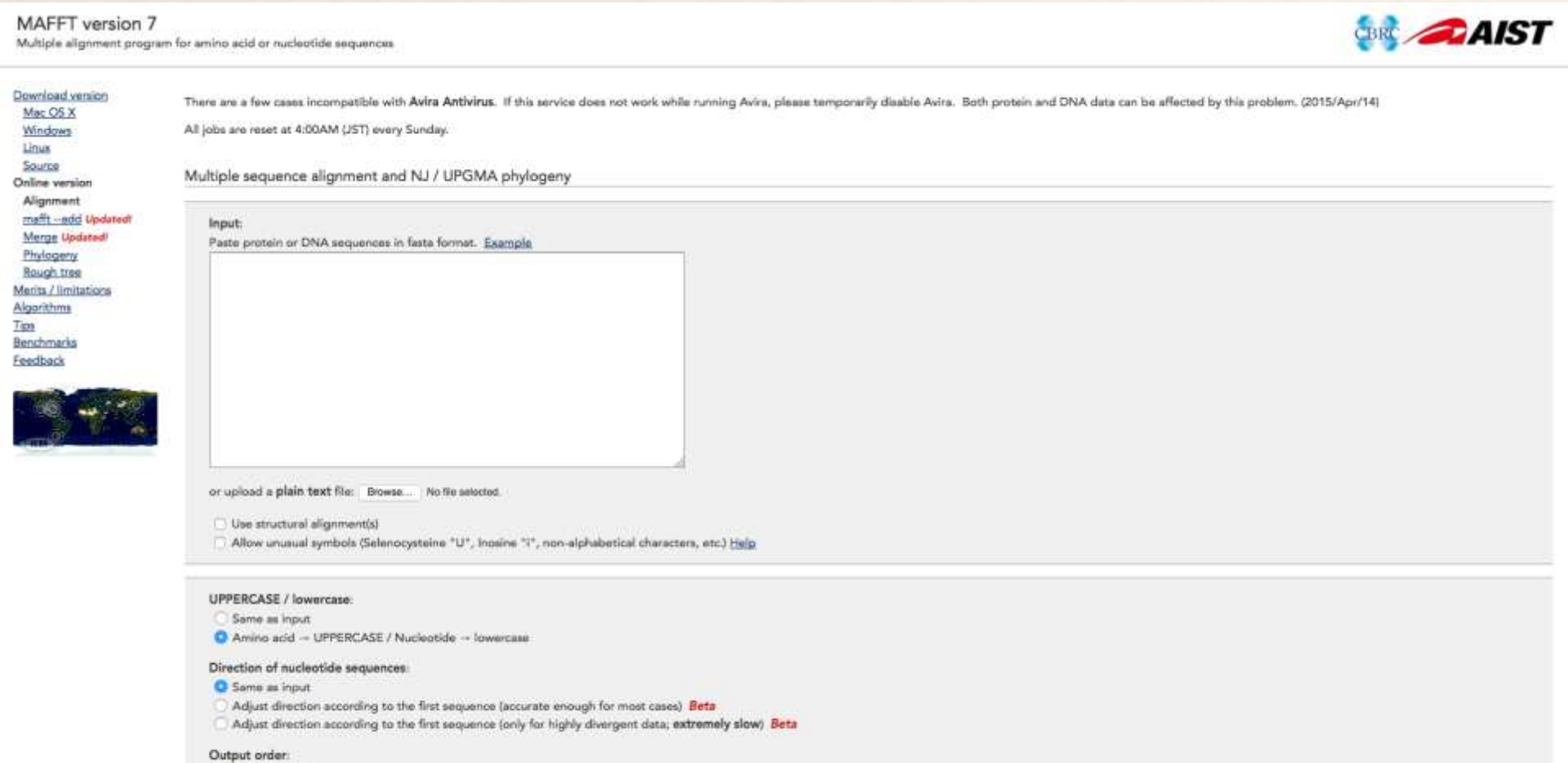
  omictools.com/orthologous-groups-c419-p1.html

- Check sequences (see Alignment)

Sequence alignment

List of alignment software: *en.wikipedia.org/wiki/List_of_sequence_alignment_software*

MAFFT v.7 (*mafft.cbrc.jp/alignment/server/*)



Multiple sequence alignme

Input:

Paste protein or DNA seque

or upload a **plain text** file:

Use structural alignmen

Allow unusual symbols (

UPPERCASE / lowercase:

Same as input

Amino acid → UPPERC

Direction of nucleotide seq

Same as input

Adjust direction accord

Adjust direction accord

Output order:

Same as input

Aligned

Notify when finished (optic

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

Advanced settings

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: BLOSUM62

Scoring matrix for nucleotide sequences: 200PAM / κ=2

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

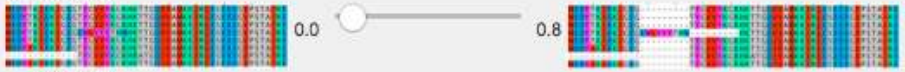
Gap opening penalty: 1.53 (1.0 - 3.0)

Offset value: 0.0 (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:



↑ Default

- (•) 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: 50 (5 - 200)

Threshold: *E* = 1e-10 (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: score=39 (E=8.4e-11)

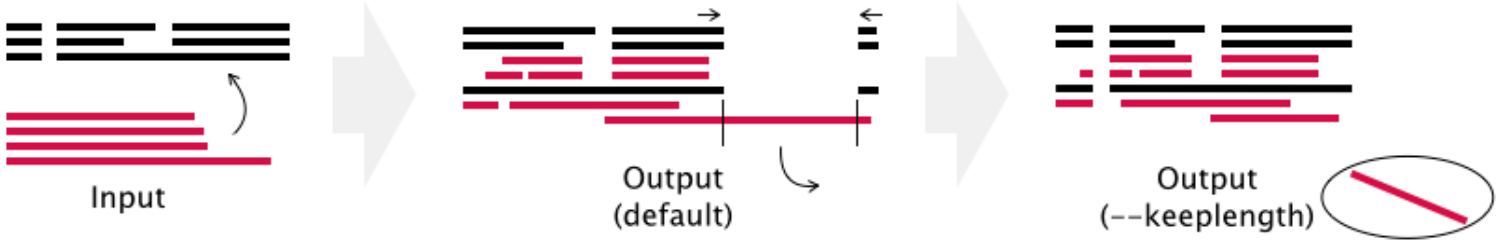
Submit Reset

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

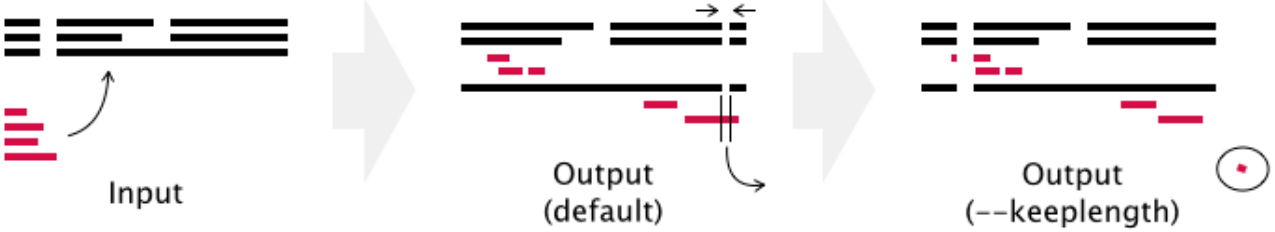
--add

Align **full length sequences** to **an MSA**



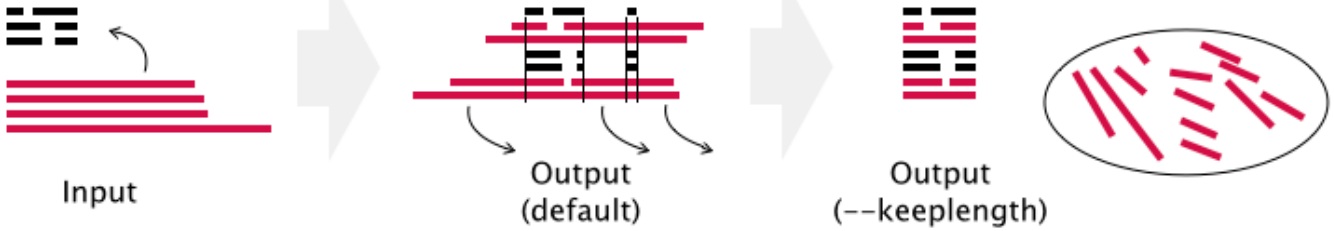
--addfragments

Align **fragment sequences** to **an MSA**



--addlong (experimental)

Align **long sequences** to **a short MSA**

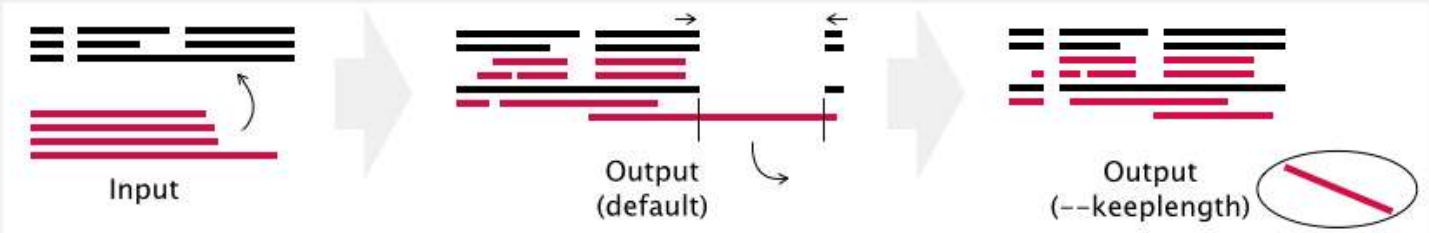


Add unaligned sequence(s) to an **existing alignment** Help

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

☐ Yes

With this option, insertions in the new sequenes 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: Browse... 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: Browse... No file selected. Clear

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

Strategy:

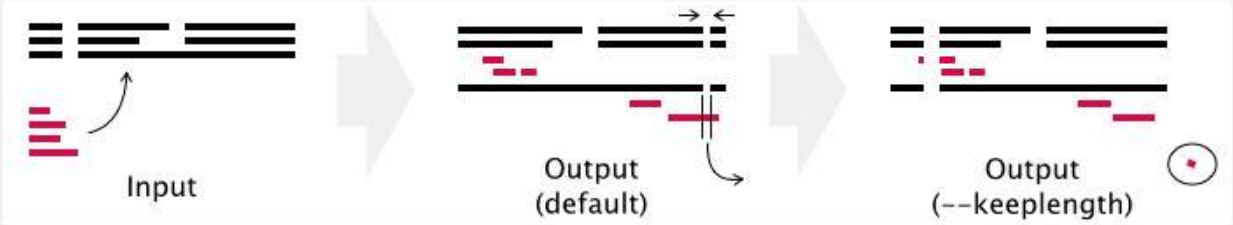
- ◉ Auto (FFT-NS-1 or L-INS-i; depends on data size)
- ○ FFT-NS-1 (Fast)
- ○ G-INS-1 (Slow; uses all-pair global alignment)
- ○ L-INS-1 (Slow; uses all-pair local alignment)

Add fragmentary sequence(s) to **existing alignment** or **sequence** Help

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

☐ Yes

With this option, insertions at the fragmentary sequenes 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: Browse... No file selected. Clear

Strategy:

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

Fragmentary sequence(s) to be added to the above alignment: Example

Gaps (if any) will be removed.

or upload a plain text file: Browse... No file selected. Clear

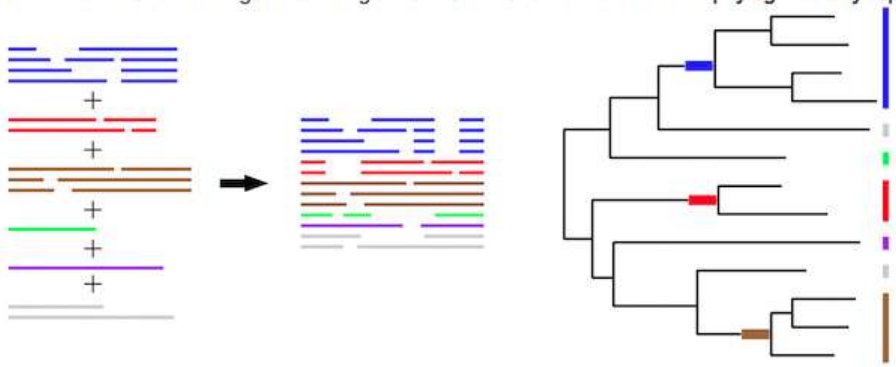
☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

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: Browse... No file selected. Clear

Sub MSA b: Example

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

or upload a plain text file: Browse... 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 do
- ( ) 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

Parameters:

Scoring matrix for amino acid sequences: BLOSUM62

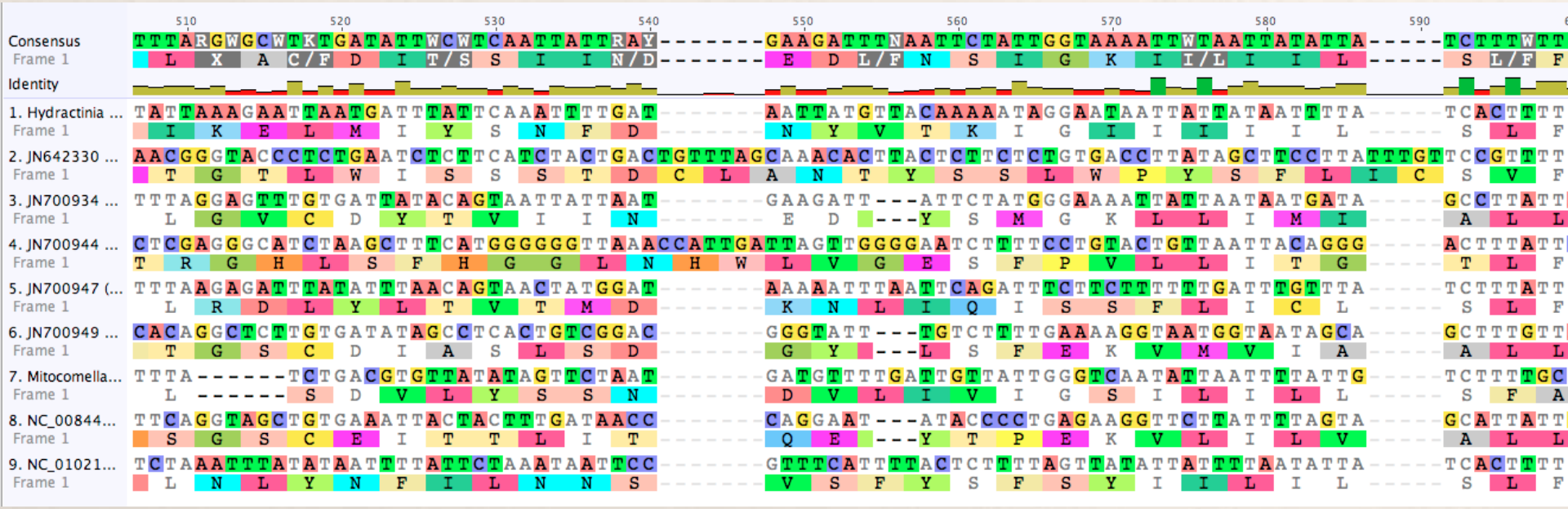
Scoring matrix for nucleotide sequences: 200PAM / κ=2

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

Gap opening penalty: 1.53 (1.0 - 3.0)

Offset value: 0.0 (0.0 - 1.0)

Protein genes have to be kept in frame!



INPUT FILE 1

Put your multiple sequence al

- The server automatical
- In the CLUSTAL form
- How to specify stop co

or

upload

Browse... nad2 translation alignm

INPUT FILE 2

Put your DNA (or mRNA) se

or

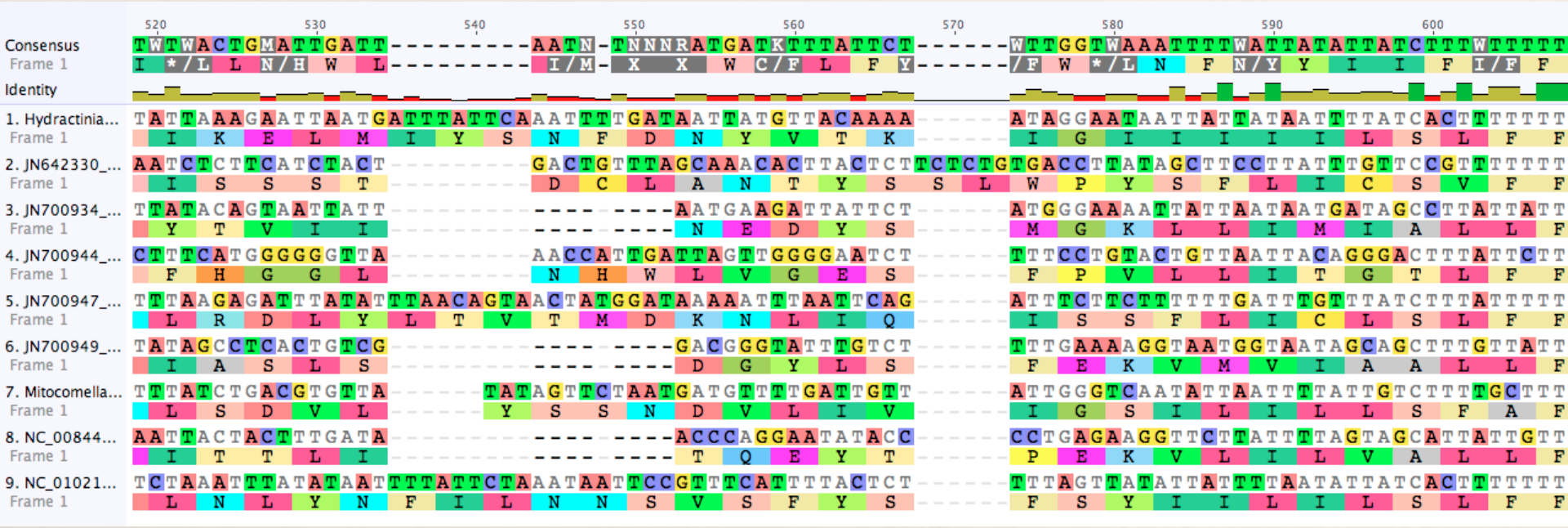
upload

Browse... nad2.fasta

Option setting

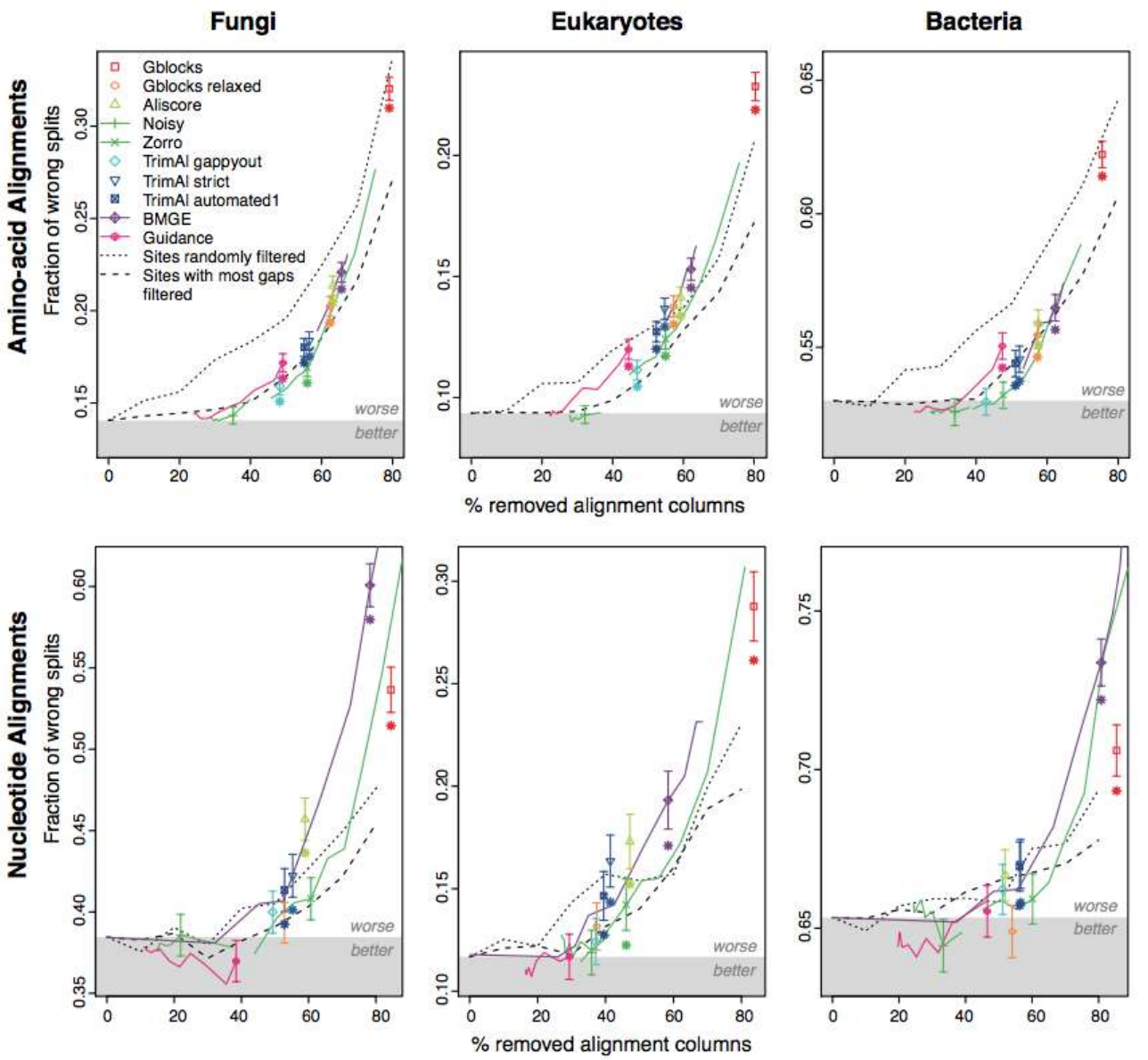
1. Codon table:
   - ○ Universal code (NCBI: transl_table=1)
   - ○ Vertebrate mitochondrial code (NCBI: transl_table=2)
   - ○ Yeast mitochondrial code (NCBI: transl_table=3)
   - ◉ Mold, Protozoan, and Coelenterate Mitochondrial code and Mycoplasma/Spiroplasma code (NCBI: transl_table=4)
   - ○ Invertebrate 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



Direct nt alignment (MAFFT-Auto)

PAL2NAL converted alignment



Gblocks

TrimAl

Noisy

Aliscore

BMGE

Zorro

Guidance

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.

Models of sequence evolution

Published models:

✧ jmodeltest v.2 (*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/*) 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

en.wikipedia.org/wiki/List_of_phylogenetics_software

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

RAxML v.8+

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.bestrtee

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.bestrtee –z raxml.boot –n raxml.run

MrBayes v.3.2

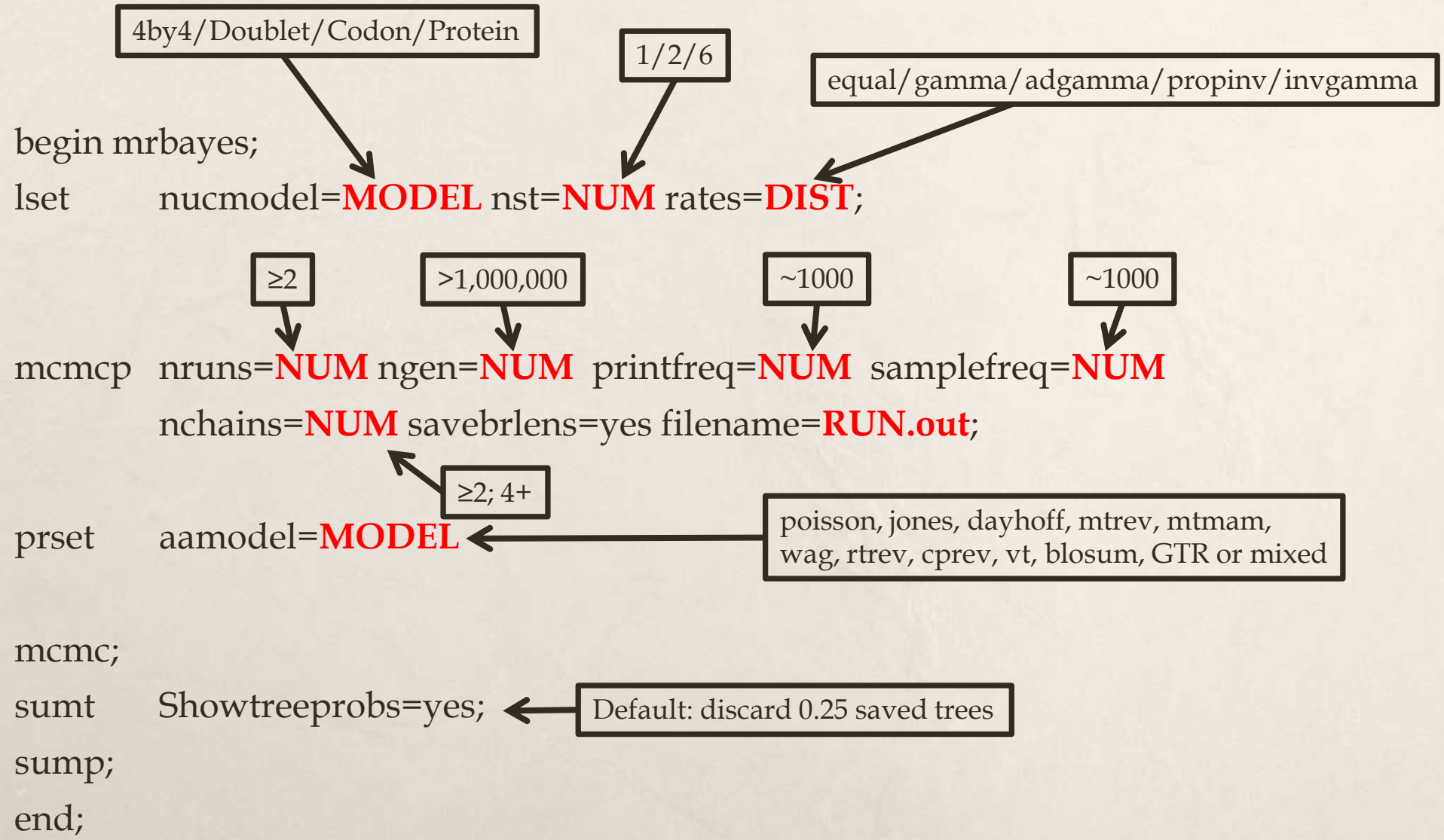
mrbayes.sourceforge.net/mb3.2_manual.pdf

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      AGTAACGGCGAGTGAACTTATAG-GA-T-AAGTACTGTGAAGGAATACCTTTTGTATAAT
        Hydractinia_symbiolongicarpus_SRR1174275_SRR1174698     AGTATTGGCAAAAGAACTTTTAA-TT-A-TAGTACTGTGAAGGAATACCTTTTGTATAAT
        Hydractinia_polyclina_SRR923509 AGTATTGGCAAAAGAACTTTTAA-TT-A-TAGTACTGTGAAGGAATACCTTTTGTATAAT

        Physalia_ESTs_Angel     CGTAACATAGTGGGGGG----------------
        Physalia_physalia_SRR871528     CGTAACATAGTGGGGGGAAGGGAACTTCTCCCTG
        Rhizophysa_DLSI230      CGTAACATAGTGGGGGGAAGGGAACTTCTCCCTG
        Hydractinia_symbiolongicarpus_SRR1174275_SRR1174698     CGTAACATAGTGGGAGAAAGGGAACTTTTTCCTG
        Hydractinia_polyclina_SRR923509 CGTAACATAGTGGGRGAAAGGGAACTTTTTCCTG

;
end;

begin mrbayes;
lset nucmodel=4by4 nst=6 ploidy=haploid rates=invgamma;
mcmcp nruns=2 ngen=5000000 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;
mcmcp 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);
mcmcp 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

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; -cat, -gam or –rr (fixed)

❖ Rates:

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

Interpreting phylogenetic results

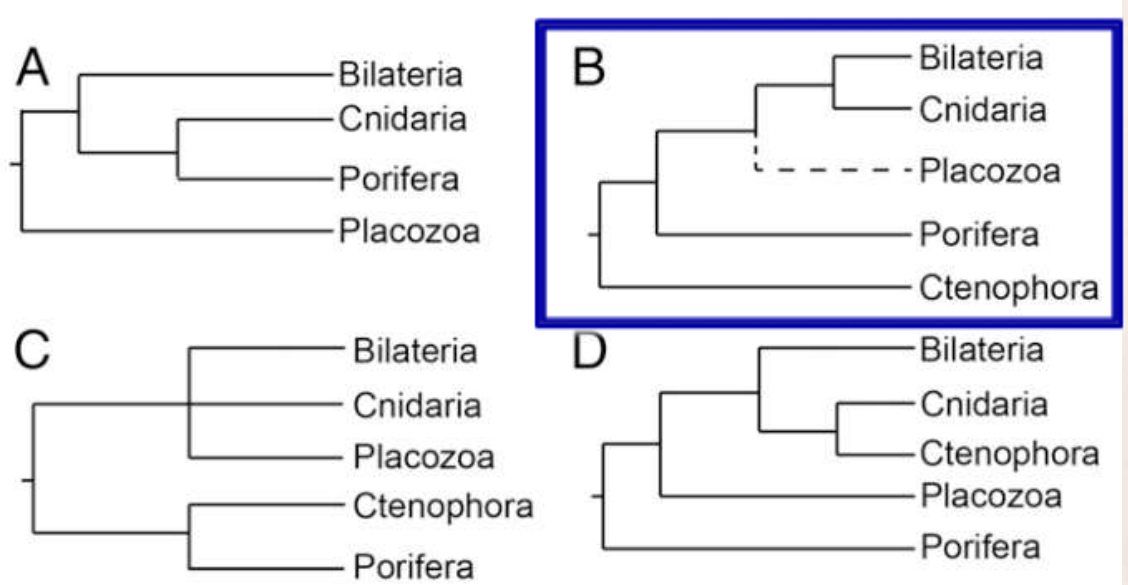


Fig. 1. Phylogenetic hypotheses from previous molecular studies. (*A*) Placozoa-sister hypothesis (2). (*B*) Ctenophora-sister hypothesis (3–7). Placozoa was not included in some studies that found support for this topology. (*C*) Ctenophora + Porifera-sister hypothesis (6). (*D*) Traditional Porifera-sister hypothesis (7–10).

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

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.

OPEN ACCESS Freely available online

S BIOLOGY

Perspective

Resolving Difficul
Sequences Are N
ore

Hervé Philippe[1]*, Henner Brink
Gert Wörheide[5,6], Denis Baurai
Manuel[4],

1 Département de Biochimie, Centre Robert-Cederg
Iowa State University, Ames, Iowa, United States of
7138 "Systématique, Adaptation, Evolution" UPMC
München, München, Germany, 6 GeoBio-Center,
Veterinary Medicine, University of Liège, Liège, Bel
ismal Biology,
é Paris 6, UMR
ns-Universität
nd Faculty of

In the quest to reconstruct the Tree o
Life, researchers have increasingly turnec
to phylogenomics, the inference of phylo
genetic relationships using genome-scal
data (Box 1). Mesmerized by the sustainec
increase in sequencing throughput, many
phylogeneticists entertained the hope tha
the incongruence frequently observed i
studies using single or a few genes [1
would come to an end with the generatio
of large multigene datasets. Yet, as so ofte
happens, reality has turned out to be fa
more complex, as three recent large-scal
analyses, one published in *PLoS Biolog*

nisms 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?