Multiple Sequence Alignment

Rosa Fernández, Lisa Pokorny & Marina Marcet-Houben



The true multiple alignment

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree

Fig. 9.1. Warnow. 2017. Computational Phylogenomics. An introduction to designing methods for phylogeny estimation. UIUC.

Multiple Seq Alignment (MSA)

S1

S4

Standard two-phase approach: 1st ALIGNMENT (positional homology)

- S1 = AGGCTATCACCTGACCTCCA
- S2 = TAGCTATCACGACCGC
- S3 = TAGCTGACCGC
- S4 = TCACGACCGACA



Fig. 9.1. Warnow. 2017. Computational Phylogenomics. An introduction to designing methods for phylogeny estimation. UIUC.

MSA Methods (MSAMs)

- → Sum-of-Pairs Alignment (SOP)
- ➔ Tree Alignment and Generalized Tree Alignment
- ➔ Sequence Profiles
- ➔ Profile Hidden Markov Models (HMM)
- → Reference-based Alignments
- → Template-based Methods
- → Seed Alignment Methods
- ➔ Weighted-Homology Pair Methods
- ➔ Progressive Methods
- ➔ Divide-and-Conquer Methods
- → Co-estimation of Alignments and Trees
- → Structure Informed Methods, etc.

Comparing MSAMs

Tool	Options	Algorithm	Alphabet
ClustalW	Defaults	Progressive	Amino Acid
Muscle	Defaults	Progressive (iterative)	Amino Acid
MAFFT	Defaults	Progressive (iterative)	Amino Acid
ProbCons	Defaults	Consistency	Amino Acid
ProbAlign	Defaults	Consistency	Amino Acid
Mummals	Defaults	Consistency/Structure	Amino Acid
Dialign-TX	Defaults	Greedy/Progressive	Amino Acid
Prank (AA)	+F (AA)	"Phylogenetically- aware"	Amino Acid
Prank	+F -codon	"Phylogenetically- aware"	Codon
BAli-Phy	Model M0	Statistical Alignment	Codon
BAli-Phy samples	Model M0	Statistical Alignment	Codon
BAli-Phy integrated	Model M0	Statistical Alignment	Codon



Multiple Alignment by Secondary Structures



MAFFT version 7

PROBCONS



T-COFFEE SIMPLE MSA

RNA



Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences A

Multiple alignment program for amino acid or nucleotide sequences

Blackburne & Whelan. 2013. Mol. Biol. Evol. 30(3):642-653.

Mean Distance between MSAMs



Fig. 1. Blackburne & Whelan. 2013. Mol. Biol. Evol. 30(3):642-653.

Distances btw. Tree Estimates from ≠ MSAMs



Fig. 2. Blackburne & Whelan. 2013. Mol. Biol. Evol. 30(3):642-653.

Even more MSAMs comparisons



Fig. 4. Liu et al. 2013. Syst. Biol. 61(1):90-106.





Multiple Alignment by Secondary Structures







PROBCONS

Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences



PASTA

PASTA estimates alignments and ML trees from unaligned sequences using an iterative approach. In each iteration, it first estimates a multiple sequence alignment using the current tree as a guide and then estimates a ML tree on (a masked version of) the alignment. By default, PASTA performs 3 iterations, but a host of options enable changing that behavior. In each iteration, a divide-and-conquer strategy is used for estimating the alignment. The set of sequences is divided into smaller subsets, each of which is aligned using an external alignment tool (default is MAFFT). These subset alignments are then pairwise merged (by default using Opal) and finally the pairwise merged alignments are merged into a final alignment using a transitivity merge technique. The division of the dataset into smaller subsets and selecting which alignments should be pairwise merged is guided by the tree from the previous iteration. The first step therefore needs an initial tree.

Acknowledgment: The current **PASTA** code is heavily based on the **SATé** code developed by Mark Holder's group at KU.



Fig. 1. Mirarab et al. 2015. J. Comp. Biol. 22(5):377-386.





Fig. 2. Mirarab et al. 2015. J. Comp. Biol. 22(5):377-386.

Running PASTA (from Command-line)

If your installation is successful, you should be able to run **PASTA** by running the following command from any location. Open up a terminal window and type:

run_pasta.py --help

Running PASTA with the --help option produces the list of options available in PASTA. PASTA automatically picks its algorithmic settings based on your input, so you can ignore most of these options (but -d is essential if you have anything other than DNA sequences). The basic command-line usage you need to know is:

run_pasta.py -i input_fasta_file

Running PASTA (from Command-line)

The -i option is used to specify the input sequence file. The input file needs to be in the relaxed FASTA format. This command will start **PASTA** and will run it on your input file.

For a test run, use the cd command to go to the data directory under your PASTA installation directory. From there, run

run_pasta.py -i small.fasta

This will start **PASTA** and will finish quickly (30 seconds to 5 minutes based on your machine). Read **PASTA** output and make sure it finishes without producing any errors. If **PASTA** runs successfully, it produces a multiple sequence alignment and a tree, which we will explore in the next step.

Inspecting the Output of PASTA

The two main outputs of PASTA are an alignment and a tree. The tree is saved in a file called [jobname].tre and the alignment file is named [jobname].marker001.small.aln. The [jobname] is a prefix which is by default set to pastajob, but can be changed by the user (see option -j below). When you start PASTA, if your output directory (which is by default where your input sequences are) already contains some files with the pastajob prefix, then the pastajob1 prefix is used, and if that exists, pastajob2 is used, and so forth. Thus the existing files are never overwritten. The name of your job and therefore the prefix used for output files can be controlled using the - j argument for command-line or the "Job Name" field on the GUI.

Tree viewing software \rightarrow <u>https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software</u>

Alignment viewing software, e.g., <u>http://doua.prabi.fr/software/seaview</u>

Running PASTA (from Command-line)

You can script a while loop in bash to run PASTA on multiple fasta files. First open a text editor

nano pasta_loop.sh

Write your bash script

```
#!/bin/bash
while read targetname;
    do
        python ABSOLUTE_PATH_HERE/run_pasta.py -i "$targetname".fasta -j $targetname
        done < targetlist.txt</pre>
```

Close CTRL+x and save your script. This script assumes all target files are in the same folder in fasta format. It also assumes that folder contains a text file listing all targets. From there, run

bash pasta_loop.sh

The command line allows you to alter the behavior of the algorithm using a variety of configuration options. Running **PASTA** with the -h option lists all the options that can be provided to the command-line (see below for the most important ones). In addition to the command-line itself, **PASTA** can read the options from one or more configuration files. The configuration files have the following format:

```
[commandline]
option-name = value
[sate]
```

option-name = value

Note that as mentioned before, with every run, PASTA saves the configuration file for that run as a temporary file called [jobname]_temp_pasta_config.txt in your output directory. You can view one of these files in a Text editor for better understanding the format of the configuration file.

https://github.com/smirarab/pasta/blob/master/pasta-doc/pasta-tutorial.md

PASTA can read multiple configuration. Configuration files are read in the order they occur as arguments (with values in later files replacing previously read values). Options specified in the command line are read last. Thus these values "overwrite" any settings from the configuration files.

The following is a list of important options used by **PASTA**. Note that by default **PASTA** picks these parameters for you, and thus you might not need to ever change these (with the important exception of the -d option):

- Initial tree: As mentioned before, PASTA needs an initial tree for doing the first round of the alignment. Here is how the initial tree is picked.
 - \circ If a starting tree is provided using the -t option, then that tree is used.

run_pasta.py -i small.fasta -t small.tree

PASTA can read multiple configuration. Configuration files are read in the order they occur as arguments (with values in later files replacing previously read values). Options specified in the command line are read last. Thus these values "overwrite" any settings from the configuration files.

The following is a list of important options used by **PASTA**. Note that by default **PASTA** picks these parameters for you, and thus you might not need to ever change these (with the important exception of the -d option):

- Initial tree: As mentioned before, PASTA needs an initial tree for doing the first round of the alignment. Here is how the initial tree is picked.
 - \circ If a starting tree is provided using the -t option, then that tree is used.
 - If the input sequence file is already aligned and --aligned option is provided, then PASTA computes a ML tree on the input alignment and uses that as the starting tree.
 - If the input sequences are not aligned (or if they are aligned and --aligned is not given), PASTA uses the following procedure for estimating the starting alignment and tree. It 1) randomly selects a subset of 100 sequences, 2) estimates an alignment on the subset using the subset alignment tool (default MAFFT-I-insi), 3) builds a HMMER model on this "backbone" alignment, 4) uses hmmalign to align the remaining sequences into the backbone alignment, 5) runs FastTree on the alignment obtained in the previous step.

https://github.com/smirarab/pasta/blob/master/pasta-doc/pasta-tutorial.md

 Data type: PASTA does not automatically detect your data type. Unless your data is DNA, you need to set the data type using -a command. Your options are DNA, RNA, and PROTEIN.

run pasta.py -i BBA0067-half.input.fasta -t BBA0067-half.startingtree.tre -d PROTEIN

- Data type: PASTA does not automatically detect your data type. Unless your data is DNA, you need to set the data type using -d command. Your options are DNA, RNA, and PROTEIN.
- Tree estimation tool: the default tool used for estimating the phylogenetic tree in PASTA is FastTree. The only other option currently available is RAxML. You can set the tree estimator to RAxML using the --tree-estimator option. However, Be aware that RAxML takes much longer than FastTree. If you really want to have a RAxML tree, we suggest obtaining one by running it on the final PASTA alignment. You can change the model used by FastTree (default: -nt -gtr -gamma for nt and -wag -gamma for aa) or RAxML (default GTRGAMMA for nt and PROTWAGCAT for AA) by updating the [model] parameter under [FastTree] or [RAxML] header in the input configuration file. The model cannot be currently updated in the command line directly as an option.
- Subset alignment tool: the default tool used for aligning subsets is MAFFT, but you can change it using the --aligner option. We strongly suggest alignment subset size should always be no more than 200 sequences, because for subsets that are larger than 200, the most accurate version of MAFFT (-linsi) is not used.
- Pairwise merge tool: the default merger too is Opal. You can change it using --merger option. If you have trouble with Opal (java version, memory, etc.) using Muscle should solve your problem and in our experience, it doesn't really affect the accuracy by a large margin.

• CPUs: PASTA tries to use all the available cpus by default. You can use --num_cpus to adjust the number of threads used.

run_pasta.py -i small.fasta --num_cpus 1

- CPUs: PASTA tries to use all the available cpus by default. You can use --num_cpus to adjust the number of threads used.
- Number of iterations: the simplest option that can be used to set the number of iterations is --iter-limit, which sets the number of iterations PASTA should run for. You can also set a time limit using --time-limit, in which case, PASTA runs until the time limit is reached, and then continues to run until the current iteration is finished, and then stops. If both options are set, PASTA stops after the first limit is reached. The remaining options for setting iteration limits are legacies of SATé and are not recommended.
- Masking: Since PASTA can produce very gappy alignments, it is a good idea to remove sites that are almost exclusively gaps before running the ML tree estimation. By default, PASTA removes sites that are more than 99.9% gaps. You can change that using the --mask-gappy-sites option. For example, using --mask-gappy-sites 10 would remove sites that are gaps for all sequences except for (at most) 10 sequences. Increasing the masking can make PASTA a bit faster and can potentially reduce the memory usage. But it could also have a small effect on the final tree. If unsure, leave the option unchanged. Note that the final alignment outputted by PASTA is NOT masked, but masked versions of the output are also saved as temporary files (see below).

Running PASTA Using Configuration Files

The configurations used for running **PASTA** are all saved to a configuration file, and also, **PASTA** can be run using a configuration file. These configuration files are useful for multiple purposes. For example, if you want to reproduce a **PASTA** run, or if you want to report the exact configurations used. Always make sure to keep the produced configuration files for future reference. Note however, that configuration files can be used as input only using command-line.

Let's open myjob_temp_pasta_config.txt under the data directory and take a look at it. Notice that the options we referred to are all mentioned here.

Now imagine that we wanted to instruct PASTA to use the JTT model instead of WAG for a protein run. Here is how we can accomplish that. Copy the myjob_temp_pasta_config.txt file as a new file (e.g. cp myjob_temp_pasta_config.txt jtt_config.txt). Then open jtt_config.txt using a text editor of your choice. Find model = -wag -gamma -fastest under the [FastTree] header. Remove the -wag option and save the config file. Note that the default model in FastTree is JTT, and therefore, when the -wag is removed, it automatically switches to using JTT. To run PASTA using this new configuration file, run:

run_pasta.py jtt_config.txt

Running PASTA Using Configuration Files

Adding custom parameters to aligners: It is also possible to add custom parameters to alignment and merge tools. To do so, you need to use the config file. Under each alignment tool in the config file, you can add an args attribute and list all the attributes you want to pass to that tool. For example, to run Mafft with your choice of gap penalty value, edit the config file under the [mafft] heading to something like:

[mafft]
path = [there will be a path here to your pasta directory]/bin/mafft
args = --op 0.2 --ep 0.2

and use this config file to run PASTA.

Note that **PASTA** does not try to understand these extra parameters you pass to external tools. It simply appends these parameters to the end of the command it executes.

Running PASTA with your own data

At this stage, if you have input files that you like to have analyzed, you know enough to start doing that.

Email: pasta-users@googlegroups.com for all issues.



Rosa Fernández, Lisa Pokorny & Marina Marcet-Houben

Multiple Sequence Alignments Trimming Multiple Sequence Alignments

-04000088889-078 <mark>R</mark> acS <u>O</u> F <mark>K</mark> LA <mark>D</mark> VTLaLaadaaaaaaaaaa		LP <mark>CHR</mark> GLLALSSPYFoorssons	HAMEAGOODEFAESCOFS	RVELRDVERACOC	양양 <u>양</u> 집 L 영 <mark>전</mark> 위 <u>신</u> 위 -
- 019960333897-039 - <mark>R</mark> = = T <mark>D & K</mark> 7 F <mark>D</mark> 17 = <mark>H</mark> = = = = = = = = = = = = = = = = = = =		FP <mark>CHR</mark> IILAAASNYFoodoodool	RISMEAG NINKES CT	KVTLF&VEPE	0 0 1 0 AA <mark>E</mark> L EF <mark>C</mark> Y
//3/03/265/8/2387		PAHRAYLAAASPYF	RAMEAGOOGL <mark>RE</mark> SCO <mark>R</mark> A	RVRLHSVPPD	· · · HLDLLL <mark>D</mark> YSY
- C1γ(G22(S1/λC2) <mark>R</mark> → A <mark>ERK</mark> 7LCVTL→ <mark>E</mark> → → → → → → → → → → → →		PPAHRAYLAAASPYF	RAMPAGODOUL RESCORA	RVRLH&VPPD	KULDLLL <mark>D</mark> FSY-
SUGGYRRSVICES ROOMERKFLOWTLEEDOUTLEEDOUTLEEDOUTLE	AAAA	PPAHRAYLAAASPYF	RAMPAGO OLRESCORA	R VR LHOVEPD	- HLDLLL <mark>D</mark> FSY
SUCCEPTION ASS ROOMERKTLDWTLLE		PRANR AVE COASPYF	RAMEAGO OLRES PA	RVRLHSVEPD	HEQUEL <mark>O</mark> FSY
SIDCOUDTRUCTS CONTINUES CONTINUES		PRANE AVERAGE PYF	RAME AR	RVRLHSVEPD	MEDILLOFSY
HODOCCUREM AND B. ARENTI DUTI F	6 065 B	FRAHRAUL COOSPYF		RCR HAVER	MEDILLOFSY
CONCERNMENT E. OFENELDUTI F.	A. 065.	FRAME OUL COOSPYF	EQMEAS	BCB HANDER	HEDLILDESY
OFWERE ARE E. AREVELOUTIE	A. 065.	TRAMP OUT ADARBYE	E AME AG OL RES PA	DUD HANDER	
	A 0.6.6	2 PAUD ON ADARDYE		BY BINAVE SA	MEDILLARSY
PERFORMANCE A CONSTRUCTION		2 PARTAYLAAAAPETEEEEEEE	DAMEAS DIDES UV	DYELUNITOD	
ADVANCES PARA AND A REPAIR AND A REPAIR AND A		2 PORTYLANA ONE			
		THE REPORT OF A CONSTRUCT			
		A PURKING A A A A A A A A A A A A A A A A A A A	HAMPAGES NLHESSEH	19 LHD VIOP CONT	
		PUTE TVLAASSAYF	HAMPAG NLHES HC	19 ELHDY I GP	····ILSELLOPUY
AMOUNTS H RENKARDY L. C		HUHRIVLAAASMYF	HAMPAO ILHES - OM	SKYULHE VOAE	···· LLOLLODFUY
MCCCC72296027 B - REBRITDYTL-C	· · · · · · · · · · · · · · · · · · ·	HURHTVLAASIYF	H SWELG LEHES 0 M	NEWDIHEWS AD	···· LLOLLODMUY
Sw00722234677 R - AECKFFD971L-C		PRCHRIVLAAASTYF	RAMFTO TLRES. UN	CARLHEARAE	····LLSLLWDPCV
DevC604222M9-302 R - 3 N KE 7 T D 9 1 L - T	· · · · · · · · · · · · · · · · · · ·	TACHR29LAA13PYF	HALLSS NMKED HE	BERTROMEPETT.	· · · · ISHMESSHAA
760020283214242 R - RAKEF DL KI-L	••••••••••••••••••••••••••••••••••••••	E EAMONALV2C3FAL	KDLYKR SSO <mark>ekr</mark> 10 + • EA	ELNEISNISAD	· · · · STEDFTERAA
Dw627243M7-349 B				ELNETRISAD	· · · · STEFFFEAA
Far906282274-628 H RAKEFIDLKI		FESHQUSLAS <mark>C</mark> BLYF	KDLVKR330EKR30··EA	ELSESSITAD · · · ·	····MTEBEIE&AA
RETRICTION REPORT AND A REPORT OF REPORT	· · · · · · · · · · · · · · · · · · ·	FESHQUYLAS <mark>C</mark> BLYF	KDLICHSVDDSSDBSNEK	ELWLSNLGAD	· · · · MULLELE
Produzation Revealed International Revealed International Revealed Strength Revealed	· · · · · · · · · · · · · · · · · · ·	<pre>>HUMPLASCSLYF</pre>	KDLIIG <mark>r</mark> avd <mark>d</mark> cande <mark>rek</mark>	- CLULSNLGAD · · · ·	• • • • MULFELE
DRACENED APPROX B - RAKED TOLE I - 9	· · · · · · · · · · · · · · · · · · ·	○F2H0旧為に管設C3FALeeeeeeeeeeeee	KD FIICHERARD SUNDO <mark>hek</mark>	- CLULSBLU.SD · · · ·	• • • • SUCTFTERSA
J99000002222-402 R - RAKAD TOLKING - STATES - STATES - STATES - STATES - RAKAD TOLKING - STATES - S	· · · · · · · · · · · · · · · · · · ·	<pre>SHOUSEASCSLYF************</pre>	KDLIERAVODSONOCREK	L'ELSELLAD · · · ·	• • • • MPEFFFF6AA
TREESERIES RATE RATER TOLER I STATES	· · · · · · · · · · · · · · · · · · ·	○■SHURSERSESEYF · · · · · · · · · · · · · · · · · · ·	KDUIERSASDSSNOCKER	ELVISBERAD · · · ·	• • • • %L <mark>E</mark> FFF <mark>E</mark> FAA.
- TARBEREN DI-AUS - H B 4 K HU L D - K I - A	· • • • • • • • • • • • • • • • • • • •	○ ■ S HORSEAS COLVENNESS	KDUICHANND <mark>D</mark> SONDO <mark>REK</mark>	ELVISBERAD · · · ·	• • • • %L <mark>E</mark> FFF <mark>5</mark> 6AA
- 2996022025/MESS R R & K AD T DL R I - &	· · · · · · · · · · · · · · · · · · ·	○ MARANESSE SEAL · · · · · · · · · · · · · · · · · · ·	KDUICRAVIDGID <mark>C</mark> CREK	ELVISBEBAD · · · ·	• • • • 3.P E P P P E A &
DMG6872S2MESS R RAKAP TOLKI. W	· · · · · · · · · · · · · · · · · · ·	○ E S H G B S L A B C S L Y F · · · · · · · · · · · · · · · · · ·	KDLICHSKODGSRCCHEK	ELVISALGAD · · · ·	• • • • 36 <mark>E</mark> EEE <mark>E</mark> PWY
- RW89105764-667 R R 4 KE 7 T DL K I - D	· · · · · · · · · · · · · · · · · · ·	○ E S H G B S L A S G S L Y F · · · · · · · · · · · · · · · · · ·	KDLIKREPRDSSRSQ-EK	ELANSNETAD	• • • • 76 <mark>E</mark> EEE <mark>E</mark> PVY.
- 8566983255-652 - R R A KE 7 T <mark>D</mark> L K I - D	· · · · · · · · · · · · · · · · · · ·	₽ <mark>₽₩</mark> ₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩	KDUIKBSSRDSAKNG-EK	ELTRISHINAE · · · ·	76 <mark>E</mark> EEE <mark>E</mark> FVY
- XECGV83229-563 - 🖪 🖪 A KE 7 T 🖸 L 🥂 I - I	· · · · · · · · · · · · · · · · · · ·	₽ <mark>₽₩</mark> ₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩	KDLIKESSRDSAKNG-EK	ELTRISHINAE · · · ·	···· WE <mark>E</mark> LEL <mark>E</mark> RVY
23*667832849-293				MSNINA <mark>E</mark>	VE <mark>E</mark> EEE <mark>E</mark> FVV.
#9453020056V-f18			n n METS n n n <mark>D</mark> BIBE <mark>C</mark> n n <mark>K</mark> Q	E TUMBON D PS	····AL <mark>E</mark> ALINEAY.
- 89/03389668/9378 <mark>B</mark> aa <mark>R</mark> 26 <mark>K L C D</mark> V T L a <u>r</u> aan ahaan ahaan a		<psahrivlaasipyfooroocol< td=""><td>HAMETSIA A DOUDEC A KQ</td><td>2E 1 171012 Q (N D 2 S r r r r r</td><td>···AL<mark>E</mark>ALINEAΥ.</td></psahrivlaasipyfooroocol<>	HAMETSIA A DOUDEC A KQ	2E 1 171012 Q (N D 2 S r r r r r	···AL <mark>E</mark> ALINEAΥ.
- #2533322006349387 F = = R 2 9 K C D 9 T L = K = = = = = = = = = = = = = =		SAHRIYLAASIPYFoodooolool	HAMETROCCORDUCTOR	2E 1710112/00 N D 2 S c 2 c c -	···AL <mark>E</mark> δLINFAΥ.
- #255332000379-383 🖪 = = 🖪 3 9 🔣 L <mark>C D</mark> V T L = 🔟 = = = = = = = = = = = = =	· • • • • • • • • • • • • • • • • • • •	<pre>\$PSAHR \$PLAAS PYFoococococo </pre>	HAMETSIA A DOUDEC A KQ	2 E 1 1010 12 0 M D 2 S	• • • • AL <mark>E</mark> ALINEAY
			o o MIFITINIO O <mark>D</mark> DIDI <mark>≣ C</mark> iolo <mark>K</mark> Q	E I VMDOND PS	• • • • AL <mark>E</mark> ALINEAY
B12662518341248 1000000000000000000000000000000000000			o o METNO O O <mark>D</mark> OIDE <mark>EC</mark> o o <u>K</u> Q	> EIVM2.0-N <mark>0</mark> 7 S → → →	• • • • AL <mark>E</mark> ALINEAY -
- 04%60%9033%4073 <mark>R</mark> ac <mark>R</mark> 20 <mark>K C D</mark> V T L = <mark>K</mark> and a constant of a constant of a		(FSA <mark>HR</mark> IVLAASIPYFooloocool	H AMIE T SID DIE <mark>D</mark> MICH <mark>E C</mark> die <mark>K</mark> Q	>EIWM201X078000	• • • • AL <mark>E</mark> ALINEAY
- #200023018/1233 B R 29 K L C D V T L . K		(FSA <mark>HR</mark> IVLAASIPYFoolooloo)	HAMETROODEDIDECOR	E 101kl 2045 0 7 8 4 4 4 4	• • • • AL <mark>E</mark> ALINEAY
1011111111111111111111111111111111111			o o MIE TINIO O O DIDIDI E CICIO KO	E VMD 0 b/ 0 7 S + + + + +	• • • • AL <mark>E</mark> ALINEAY -
Hts/002/2099/1-809			olo MFTN olo o <mark>d</mark> bib <mark>i≣O</mark> olo <mark>R</mark> Q	EIVMDOND PS	· · · · AL <mark>E</mark> ALIBEAY
23%00/5932/1.7/8 LRHKQL <mark>CD</mark> VIL. <u> </u>	<u>.</u> <u></u>	IPAHRLYLSSVS <mark>D</mark> yfoddododd	A AMETSON DVREAD KO	E E I <mark>K</mark> ME & V E P N & 5 & 5 &	ALWALVOYSY
2000/0030/1607 RRPSNCDIENLPL-OFMAS95ELPVA53EF	FTGTSDOAPLLOKODOEDGVNEEDHA. SE	335 <mark>CR</mark> lvLSSVSDYF	AAMETS	EET KME XVE ZNAAAAA	ALWALVOYSY
ASS20006069/1708 L RHK01CDVIL V	AGO 8	IPAHRLYLSSVS <mark>D</mark> YF	AAMETN DVREA RO	EIKKESVERN	SLWSLIGYAY
AGC0018628/1-591 L RHK01CDVIL V		I PAHRLYLSAVSDYF	AAMETN DVREA RO	EIKKESVEPN	3LWALVOYAY
AKA09918825/1-709 L RHKOLCOVIL	VAODES	IPAHRLYLSAVSDYF	A AMETR DVREA	EIKNE 2 YE 2 N	BLWALVEYAY
MAC 2018227/1848 HOLD REEVOD 2TSEE		TSSCRLVLSAVSDVF	QAMETN DVR RQ	EIKKESYEPH	SLWALVOYAY
DevCE07430/1-518		HKSILVLSSVSDYF	AAMETS DYREA KO	EIKME OND PD	ALWYLYBYAY

Are these sites informative?

	_		1			 L .				
Discoversion R - SQ PKLAD	TLIL		066 - 86	. P C	RALLALSSPY?	 HADFAG -	0 0 0 7 0 E S	FSARVE	1 RDVEPA	 VVGQLV <mark>D</mark> FVY
02:000093399/3-097 R T D G K F F D I	VI.E		DAD - ER	FPO	RITLAAASEY?	 R SHIF AG -	■ kikci≤E €	 СТРКУТ 	LFGVEFE	 AAETLIEFCY.
10/000000/9/1-037 R A ERM F L D1	TLIE		AGG - RD	FPA	RAVLAGASEYS	 RAMEAS -	OLRES	BAERVE	LHOVEPD	 MLOLLL <mark>D</mark> YSY
OVER 28 (BULLER R	TLIE		LOG. RD	FFA	RAVLAGASEVE	 RAMEAS -	OLRES	BAERNE	LHOVEPD	 MEDILLERESY
RUCCHTTCHLER R DERMELDI	TL.E		Dee. RD	FPA	RAVLASSERVE	 ROMESS.	OL RES	BOERKE	LHOVERD	 MEDILLEREN
RUPESSTONARD R OFPICEID	TL		See. Pr	FRA	RANIDSAREVE	 ROMESS.	OL RES	RAFRER	HOUPPO	 NUMBERY
Nucreaseses			Lee. Pr	E P A	RANI DESCRIVI	PONTES.	OT PES	P SERVER	HOUDD	NIGHTERSY
Hermonial strate	11		Geo. Pr	1.0.0	RAVI DESCRYT	PART SO.	A DE RES		HOUPP	MIGHTINPRY
Manager Strategy and Day 1 and Day 1 and			C		DAVI CEADEVI	DAME NO.			L L CALER	 MERICIPERS
			Lee at		BANI SEACOPT	BARE SO.			LONDO	NERREPERT
					DANI ALARDES	 PART AC				 MERICIA PERM
AND			COOL NO.		PROLOCIES STORE	 PROPERTY AND				
VINCENSELLING R REAR FOR	14.8		1000.000		REPERMENT	 FOOD DOS -	S S N S S	0 0 📕 A 🔽 🗆 O 🗖	THNALAN	 U LE SELEE <mark>M</mark> S <mark>e</mark> X
ADDRESSIVEREZ R O ALE CIRFFOR	14.8		1000.000		REVERGENTS	 RAD COC				
THORE AND A PROPERTY OF A PROP	14.8		000.000	1 1 1	REPLACENTS	 RUD DS -	· · · · · · · · · · ·	이이에운영감이는	THOMIGH	 ULSLLL <mark>DFC</mark> Y
ZERREFTS/1677 R REQUEFTDI	TV-S		766-3 0	F P 91	REVEAGASNYS	 RADIFAG	B-RES	· · HYDINE	THDALAS	 IL SLLL <mark>DFC</mark> Y
ZEROREFERST R ALCHEFFOI	TV-S	• • • • • • • • • • • • • • • • • • • •	Nee - 20	F P O	RTYLAGASNYS	 RAME AS -	· · <u>R</u> - RES	· · HYDDYE	LHDVT02	 IL SLLL <mark>DFC</mark> Y
DWOOTSTOSTOSTS R AERKFFD	TLIC		REGARE	HC	RTVLAAASNVF	 RAMEAD	TL <mark>RE</mark> S	VM <mark>DR</mark> VV	LHEVERE	 LLOLLV <mark>d</mark> e <mark>c</mark> y
76/0025123/3-587 R AERMEED	TLOC		REGARD	HC	RTVLAAASTVF	 RAMPTS -	· · TL <mark>RE</mark> S	VM <mark>DR</mark> VU	LHEVERE	 LLOLLV <mark>D</mark> F <mark>C</mark> V
Sw00352239-911 R · · A B OK F F D	TL · 🕻	• • • • • • • • • • • • • • • • • • • •	E C + RE	H C	RTYLAAASTYS	 RAMP TO -	· · TL <mark>RE</mark> S	• • <u>Madc</u> an	LHEVSAE	 LTOTIADECA.
DPHORE499495-PER R · · ENKELTE	IL T			1 4 2	RVYLAANSPYT	 HALL33.	<u>98'KE</u> O	• • <u>REORS</u> H	TROHEF	 IN HOLVRHAY
REGOODERASHGAD R R A KEEFT BI	X 1+1		S KI		DRYLAG <mark>C</mark> SLYC	 KD CAKES	SC CKRNC	• • 🚺 AL 🖬 L B	MSNISAD	 AFLUEFCAA.
29:667253125-546 R · · · · · · · · · · · · ·				1.1.1		 	<u>ETR</u> SG	• • <mark>EK</mark> L E LS	NTNESR	 VE <mark>E</mark> EEE <mark>E</mark> VV
AU0002623/9-628 R R A KE F T D (8141		- 10 E G KE	F V	DINULAS <mark>C</mark> SLYI	 KDLWKRS	56 <mark>EKR</mark> 86	··EALELN	MSNITR	 VE <mark>CLE</mark> RVY
#2528328305/7-633 R 🛶 R A K A F T 🚺	81 · V		DEGORE	F V	DNVLAS <mark>c</mark> slyf	 KOLIDES	VC 0 3 5 C 5	S REKLELU	LSNLCAD	 VEELLEFVY
RE0660453278-660 R RAKAFT D	RIN		DEG RE	FEV	DNVLAS <mark>C</mark> SLY7	 KOLIORS	VC D6820	E REKLELV	LSNLCAD	 VE <mark>E</mark> LEE <mark>E</mark> FVY
HOLDONGE20/1-354 R R AKAP T D L	KI-V		VEG RE	F V	DNVLAS <mark>C</mark> SLY?	 KOLIORS	VC 0 3 6 2 6 1	REKLELV	LSNLCAD	 VEELEEFVY
H:10016628/14782 R R AKAF T D .	K 1. V		REG-RE	FEV	DIVLAS <mark>C</mark> SLY7	 KDLIDRS	VC <mark>D</mark> 36Q94	REKLELV	LSNLOAD	 VE <mark>E</mark> LEEEEVY
PROCESSANGES R BAKAFT DI	KI-V		EG-RE	FEV	DNYLAS <mark>C</mark> SLYS	 KDLICRS	NC <mark>D</mark> 36084	REKLELY	LSNLOAD	 VEELEEFVY
FUCCESSIVERS R RAMARTON	×1-1		SEG-RE	F V	BNYLAS <mark>C</mark> SLYF	 KDLIDRS	VC <mark>0</mark> 3000	PREKLELY	LISNLOAD	 VLELLLERVY.
SHACE22021/MEET R BY KAFT 01	×1.9		SEO. RE	E V	ONVLAS <mark>C</mark> SLV3	 RDLIDRS	NCDOSOC	2 BEKLELY	LONLOAD	 YLELLLERYY.
CMACCEP209PAGEE R BAKAFTDI	×1.9		EG. RE	E E VI	DNYLAS <mark>C</mark> SLYS	 RELIDES	NC BASS	REKLELY	LONLOAD	 VLELLLERVY
Generation Contraction Review	1.1		E CO. KE	1 V.	RIVLAS <mark>C</mark> SLYE	 KOLIKRS	PROBAT	- EKLELA	MOBLER.	 VERLEEPVY
XHOOMING CHARGE IN A REAL FOR THE	×1.1		C. G. KI		REVLAS <mark>C</mark> SLVI	 KDLIKRS	C B D C A K D		KISBLER A	 ST FLUE FRY
2HOOM 22H AGO R R SKEPT D	R 1 - 1		E G . KE		DEVLASESLY:	 KOLIKES	SBDGARD		NSBIDE	 VEFELLERWY
NHOUSE 2284-363	F			- 1		 			PERIOR	 WELLERWY.
A0xx883303423-003						 - HETEL	DUVEC		NDSNDES	 ALEALINEAY.
1810 1930 1930 5/3 771 B	T	RVI a	CO HE	5.5.6	RIVI 668 FVS	 HARTEN	DUVEC	DEL1	NDSNDFS	 ALEAL DEAY
A001110005223.537 B	TI	M.C.	i a h	1.5.6	BINIAASIEVE	 HARTE	DUVEC	NODE IN	HD SIX D F R	 AL FOLL NEAM
ASU 1920052/1.522 R	TI		ILG . HE	1.4	REMEASERVE	HOMETH	DWYEC	DDELD	MOSNERS	ALE OF THE AV
05/001100500 (1000 m 1 1 m 0 m 0 m 0 m 0 m 0 m 0 m 0 m				1.0		HZ TH	NUVE C	DONE 11	MIDANDER	ALCOLUME AV
21-2020-23-041-040						- HETH	DUUE C		140 5 4 5 5 5 5	 ALCOLINE ON
						 LOBIE TR	a subsection		100 000 0 0 0	 ALCOL INFOM
					REPEARDER'S	 CAMP THE			MOONDOO	 ALCOLINEAT
Manager and the	1		1.0	2.4	RISLARDIPYD	 NET DI -	· · · · · · · · · · · · · · · · · · ·		NOCHERS	 ALCALINEAY
ACCURATE AND A CONTRACT OF A DECISION OF A DECISIONO OF A D						 10000	· · · · · · · · · · · · · · · · · · ·		KIN ON D P S	 ALCOLUMN AND
	1.1.1					 · • REV T R •			CONCERNE.	 WE BELLINE HY
			Regione RR	1.1.1	REPERSENT	 *****************	S S S S S S S S S S S S S S S S S S S	DAEEIK	0.04.68	 WEDNE VETSY
APPEVERATION REPENDED	PL C	UASOSELPVASSEFPTOTSEGAPLLGEPDGEDGVNET		0.05	BLAFSRASDAL	 ARRETS -	BE	ZEEIK	CEONERN.	 ALOREVEYSY
4252233333333347733 L - RHK2LCD	1 L - M		149 P BB	PA	REAFSSASDAL	 CONDET N	a a PARE A	BZEEIK	KOEGANE PA	 SLOBLIGYAY
AMOBELSESSION SET L OF HIKDLCD	1 L - V		NO CORF	P A	REVESAUSDYS	 OXIOF TN -	· · PREA	OB2EEIK	이번하였던린킹	 SPORTAGUAR
1990 001 032077 709 L R H K D L C D \			PAGORE	I P.A.	REVESAUSDYS	 o aldifi T N -	D D REA	BDEEIK	MEGVEPN	 SLOALVDYAY
A\$40.001.382177-873 ₩0/L DR > E + D D	TSEE		ENE-SD	1 5 5 1	REVESAUSDYS	 AARFTN-	- PRREA	RDEEIK	MEGNEPN	 SLOALVEYAY
Diversifies		***************************************		HE	ILALSSAS <mark>D</mark> AS	 AAMETS-	• • 🛛 🖓 👧 🖻 🖓	K D E E I K	MEONEPD	 ADMIDVBYAY

While gaps represent in theory insertions and deletions, phylogenetic methods are unable to use this information to reconstruct a tree



Alignment trimming

There are multiple programs to trim a MSA

Filtering methods	Type of "undesirable" sites filtered out by the method	Accounts for tree structure?	Uses a substitution matrix or model of evolution?	Adapts parameters for particular data sets?	References
Gblocks	Gap-rich and variable sites	No	No	No	Talavera and Castresana (2007)
TrimAl	Gap-rich and variable sites	No	Yes	Yes	Capella- Gutiérrez et al. (2009)
Noisy	Homoplastic sites	In part	No	No	Dress et al. (2008)
Aliscore	Random-like sites	No	Indirectly	No	Kück et al. (2010)
BMGE	High entropy sites	No	Yes	No	Criscuolo and Gribaldo (2010)
Zorro	Sites with low posterior	Yes	Yes	No	Wu et al. (2012)
Guidance	Sites sensitive to the alignment guide tree	Yes	Indirectly	No	Penn et al. (2010)

Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference

Ge Tan; Matthieu Muffato; Christian Ledergerber; Javier Herrero; Nick Goldman; Manuel Gil; Christophe Dessimoz ☎



A tool for automated alignment trimming

http://trimal.cgenomics.org

Current stable version is 1.3 but version 1.4 contains a lot of new options so you may want to use that.

BIOINFORMATICS APPLICATIONS NOTE

2009, pages 1–2 doi:10.1093/bioinformatics/btp348

Phylogenetics

trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses

Salvador Capella-Gutiérrez, José M. Silla-Martínez and Toni Gabaldón* Comparative Genomics group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader, 88 08003 Barcelona, Spain



alignment trimming

What can trimAl do?

1.- It allows the user to trim user-defined columns or sequences

2.- It allowed the user to define some thresholds and trim the alignment according to those thresholds:

- A.- Gap thresholds
- B.- Similarity threshold
- C.- Consistency threshold Needs multiple alignments

3.- It allows the user to define a minimum percentage of alignment that has to be retained after trimming (Conservation score).

4.- It implements several automated methods that will chose the best trimming strategy based on the alignment.

Automated methods:



-strict -strictplus



Gappy-out + trimming by similarity scores \rightarrow they will only delete blocks of data so if one column has been marked to be deleted but it is surrounded by non-marked columns it will be kept in the alignment. The two methods differ on how they define the block size.
-automated1

Will use a decision tree in order to choose which of the two methods will work best on the alignment.



Additional utilities implemented in trimAl

1.- Transform your amino acid alignment to a codon based alignment.

-backtrans <inputfile> Use a Coding Sequences file to get a backtranslation for a given AA alignment

2.- You can delete a set of columns and obtain the alignment of the columns you have removed. (i.e. Keep only those columns that have at least one mutation)

-complementary Get the complementary alignment.

3.- Delete gaps

-nogaps	Remove	all positions with gaps in the alignmen	nt.
-noallgaps	Remove	columns composed only by gaps.	

http://trimal.cgenomics.org



A tool for automated alignment trimming

OVERVIEW

- Introduction
- Publications
- News
- trimAl team
- FAQ
- Mailing list

DOCUMENTATION

- Getting started with trimAl v1.2
- Use of the trimAl v1.2 command line
- Use of the trimAl v1.2 webserver interface
- Getting started with readAl v1.2

Changelog

DOWNLOADS

OLDER VERSIONS

This is trimAI's information page. Also, you can find information related to the readAI program.

trimAl is a tool for the automated removal of spurious sequences or poorly aligned regions from a multiple sequence alignment alignment

trimAl can consider several parameters, alone or in multiple combinations, in order to select the most-reliable positions in the alignment. These include the proportion of sequences with a gap, the level of residue similarity and, if several alignments for the same set of sequences are provided, the consistency level of columns among alignments. Moreover, **trimAl** is able to manually select a set of columns to be removed from the alignment.

Additionally, **trimAl** implements a series of automated algorithms that apply different thresholds, based on the characteristics of each alignment, to be used so that the signal-to-noise ratio after alignment trimming phase is optimized.

Moreover, the user can remove spurious sequences from the alignment before applying any method to improve the alignment's quality.

Among **trimAI**'s additional features, **trimAI** allows getting the complementary alignment (columns that were trimmed), to compute statistics from the alignment, to select the output file format, to get a summary of trimAI's trimming in <u>HTML</u> format, and many other options.

trimAl is being developed by the Comparative Genomics Group at the Centre for Genomic Regulation (CRG) at Barcelona, Spain.

In this site you can find the user manual, publications, news and also information related to the **trimAl** package. In this package, you can also find **readAl**, a tool for format alignment conversion.

You can also use an online version of **trimAl** and **readAl** through **Phylemom2 webserver** at the Centro de Investigacion Principe Felipe (**CIPF**) in Valencia, Spain.

Phylemon2

[[downloads]]

Login



A tool for automated alignment trimming

OVERVIEW

- Introduction
- Publications
- News
- trimAl team
- FAQ
- Mailing list

DOCUMENTATION

- Getting started with trimAl v1.2
- Use of the trimAl v1.2 command line
- Use of the trimAl v1.2 webserver interface
- Getting started with readAl v1.2
- Changelog

DOWNLOADS

OLDER VERSIONS

trimAl v1.4 (Beta version)

After the publication of a stable version of **trimAl v1.3**, **trimAl v1.4** will be our development version where we will add new functionality and continue improving trimAl's implementation. Official repository is here: **whttp://github.com/scapella/trimal**.

trimAl v1.3 (Release Candidate)

trimAl v1.3 has been our development version for almost 2 years. You can find the official repository here:
http://github.com/scapella/trimal.

Development of **trimAl v1.3** has been frozen since we expect to release a stable version during May 2011. We'll only fix bugs before releasing this version.

If you have any suggestion or idea, don't hesitate to contact us. We'll add new functionality to trimAl v1.4, our new development branch. To get the latest news, you can subscribe to our mailing list.

trimAl v1.2 (Official release)

trimAl v1.2 can function across different platforms, here you can download the versions for three different Operative Systems, if you experience problems, you do not hesitate to contact us.

You can subscribe to our mailing list to get the latest news about these programs.

- trimAl v1.2 either for Linux or MacOS
- trimAl v1.2 for Windows

Remember that **trimAl v1.2** is a program that works with command line interface. Windows' users have a compiled version of **trimAl/readAl** in the 'bin' directory. Also, these users can compile the source code with a cross platform compiler, i.e. MinGW.

downloads.txt - Last modified: 2011/04/29 14:55 by scapella

readAI: Reformatting MSAs

One of the main problems of alignments is the fact that different formats exist, and there may not be a match between the output format of an alignment program and the input format the next program needs.

#NEXUS

[!Imported PHYLIP file "030103phylip.phy" (Fri Jan 03 12:43:38 2003)]

Begin data;

Dimensions ntax=29 nchar=949;

Format datatype=nucleotide gap=- missing=? matchchar=. interleave; options gapmode=missing;

Matrix

487GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTCCATCGCCCCGGG
484GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTC-ATCGCCCCGGG
476GJS	AACGTTACCAAACTGTTGCCTCGGCGGGGAAATTCCATCGCCCCGGG
481GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTCCATCGCCCCGGG
497GJS	AACGTTACCAAACTGTTGCCTCGGCGGGGAAATCTCATCGCCCCGGG
501GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTTCAT-GCCCCGGG
477GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTCCATCGCCCCGGG
493GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTTCAT-GCCCCGGG
486GJS	AACGTTACCAAACTGTTGCCTCGGCGGGAATTCAT-GCCCCGGG
TminuEX	AACGTTACCAAACTGTTGCCTCGGCGGGAAAATTTCAT-GCCCCGGG
Tharzia	GCCCATCTACGGAAGATCATTCCAGAACACCGCTGGTATTGGCCAGACT
Tvirens	GCCCATTTATGGACGATCATTCCAGAACACCGCTGGTATTGGCCAGACT
Thamatum	GCCCATCTATGGACGATCTTTCGAGAGCACCGGTGGAATTGGCCAGTCT
473gjs	GCCTATTTACGGACGATCCTTCGAGAGCACCACTGATATTGGCAAGTYT
Hpiluli	GCCTATTTACGGACGATCTTTTGAGGCCACNACNGATATTGGCCAGTTT
153DGJS	GCCTATTTACGGACGATCTTTTTAGKGAACCACTGATATTGGTCAGTCT
130DGJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGGCT
135DGJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGGCT
139DGJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGTCT
147GJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGTCT
138DGJS	GCCTATTTACGGGCGATCTTTCGAGGGCACCACTGATATTGGCAAGGCC
491GJS	GCCTATTTACGGACGATCTTTTGAGGGCACTACTGATATTGGCAAGGCT
460GJS	GCCTATTTACGGACGATCTTTTGAGGGCACTACTGATATTGGCAAGGCT
467GJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGGCT
124DGJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGGCT
150DGJS	GCCTATTTACGGACGATCTTTTGAGGGCACCACTGATATTGGCAAGGCT
croceum	GCCTATTTACGGACGATCTTTTGAGGGCACTACTGATATTGGCAAGGCT
polysp	GCCTATTTACGGACGATCTTTTGAGGGCACTACTGATATTGGCAAGGCT
toment	gcccatctacggacgatcattccagawcaccgmcggyattggccakact
;	
End:	

begin sets:

charpartition genes = ITS:1-383, EF:384-495, ECH:496-.;

end;

TOVO	DAT	174	- 220	
PIRTZ	PCA I	124	- 239	

HFC
GGSLINDQ-WV-VSAAHCYKSRIQVRLGE-HNINVLEGN
EQFVNAAKIIKHPNFDRKT-LNNDIMLIKLS
SPVKLNARVATVALPSSCAPAGTQCLISGWGNTLSSGV
ITDNMVCVGFL
-EGG-KDSCQGDSGGPVVCNGELQGIVSWG-YGCALPDNPGVYTKVCNY
VDWI

>016LB2 AEDAE/136-374

ILNGIEADLEDFPYLGALALLDNYTSTVSYRC	
GANLISDR-FM-LTAAHCLFGKQAIHVRMGTLSLTDNPDED	
APVIIGVERVFFHRNYTRRPITRNDIALIKLN	
RTVVEDFLIPVCLYTEQNDP-LPTVPLTIAGWGGNDSAS	
UMSSSLM-KASVT-TYERDECNSLLAKKIVRLSNDQLCALGRSEF	
NDGLRNDTCVGDSGGPLELSIGR RKYIVGLTSTG - IVCGNE - F PSIYTRISQF	
IDWI	

150 1075	
Phy0007P00 GIBZE	ME-TOSSNGN-APAT-K
Phy0007T05 GIBZE	VGTSEY-AR
Phy000D2RD SCHPO	MG-K
Phy000FOYT ASPFU	MFHTAL-N-PRA-ATGEERR-G
Phy000FR4G ASPEU	LRRRLHHSKH-R-ICL-OFALSLRKTSIGSVDSMGESCKAP-0
Phy000FRUS ASPFU	MDPNSSSST-OSLPKVSLPRITTAKP-RT
Phy000FSLT ASPEU	PSPTAER-SVNSKVATPRSANPSS-WTS
Phy000FU66 ASPEU	MASHRTS-LO-PG-SSK-YINL-NERYKYAWMO-HOLOPSS-PORP-K
Phy000FUZH ASPEU	MEHTROOSVG-R
Phy000FV0G ASPEU	YOLOSGGYFHYARSP-L
Phy000FW1S ASPFU	DY-ONVRPLRR-E
Phy000FWHP ASPEU	MAS-T
Phy000FX8T EMENI	GSMSEO-RPSEPS-TPGSKIPIPRVSOLRA-YG
Phy000FXVO EMENI	PAPGAEP-S
Phy000FYYO EMENI	MFIAGRTSTG-R
Phy000FZYA EMENI	PNDIR-A-LIT-GASGGIGAACAHOLHLALTNS-AV
Phy000G015 EMENI	MSSRGPT-R
Phy000G06G EMENI	GOYEALOVSPILRORSTL-A-VIG-GDLLONGHSASEAYGSVFRSR-H
Phy000G0IY EMENI	
Phy000G1Z6 EMENI	MFHTROSSIG-R
Phy000G28K EMENI	GAOPGILHASRLOGR-K
Phy000G570 ASPNG	DISOSR-R
Phy000G8Y0 ASPNG	MYRS
Phy000G9MD ASPNG	MASA AD EDDSDSFF - E - DHD - ASPGHDLMMKDDAL GDGKP DPLPMO - K
Phy000GA9V ASPNG	MNQTHREGEPLÂ-K
Phy000GBJT ASPNG	MAS-T
Phy000GBKP ASPNG	SVPDGFPTRFNSG-I-IPSTPTKLCNNIWLEMDSSGRSRGKS-R
Phy000GEPP ASPNG	KRSEV



PHYLIP format

readAl: Reformatting MSAs

readAl is a sister program to trimAl that allows us to convert alignment between each other.

readal -in [input file] -format -out [output file]

Input file \rightarrow Alignment file

Output file \rightarrow Resulting file

Format \rightarrow Can be any of the formats that readAl has and that you wish to use as output: Fasta, phylip, mega, nexus, clustal,...

readAI: Colouring MSAs

	10	20	30 40	50	6	0 70	80	9 9	0 100	110	9 120
Phy0007P00 GIBZE			-+============+ 	E - TOSSNG	N - APAT - I	K-RRAVLACDRO	RL - K - KYK -	-C-S-EA0P-	- CSHCKR SGA	E-CKYGR-C	Y-R-PN-
Phy0007T05_GIBZE	MR	G-A-FIP-WID	KYGELLRVCG	-N	- VGTSEY - A	RPFSTVTPF	F-IS-SLST	D-LMSP-	-HAHRMRKEVY-	AH-A	K-IA-S-
Phv000D2RD_SCHP0					MG-	KPKRIPRACDMC	-KE-KIR.	CGK0A-	-CSNCVSHGI	P-CVETA-F	PKRR-GO-
Phy000F0YT ASPFU MEHT	A	L-N-PRA-A			TGEERR-	GRRPGLRACKEC	R-RR-KIRO	T-GTHP-	- CEPCLYYKKPE		-RVPH
Phy000FR4G ASPFULF	RRL HHSK	H-R-ICL-OFA	LSLRKTSIGSVD	SM	-GESCKAP-	OIIRKSFACDEC	K-RB-KIRC	CT-GGDN-	- CTNCVRD AK	P-CRYSS-F	-SOK-LL-S-
Phy000FRUS ASPFU	MDPI	NSSSST-OSLP	KVSLPRITTAK-	P-R		RRASRACISC	R-EO-KAK	C-T-GGSP-	- CERCRE LKI	E-CVEVAN-	KNEKI-
Phy000FSLI ASPFUF	S	-PTAER-SVNS	KVAIPRSANPS-	S-W	T	SSGRVSRACENC	R-EO-KAK	C-S-GHRPT	- CORCOE SGI	R-CSYGDR-	KREKM-
Phy000FU66 ASPFU MASH	RTS-LQ	-PG-SSK-YIN	L-NERYKYAWM-	Q-HQLQPS	S-PORP-I	K-QRATQACDRC	RL-K-KYK-	C-D-ELYP-	- CTHCRK NKL	E-CIYOR-N	Y-R-QRE-D-
Phy000FUZH ASPFU MFHT	E	G-P-GAS-APA	K GRE		R005VG-I	RRVTTLRACTS	B-HB-KIKO	D			
Phy000FV0G ASPFU		M	SGHPSEELSAD-	YQLQSGG	YFHYARSP-I	LPTRV-SACER	R-H-KTR-	-CD-PFR-P	C-SLCARAKV	D-COPLP-A	45-0
Phy000FW1S ASPFU			MA	DY - QN	VRPLRR-I	ERLKVSTACQAC	B-LB-KAK	CGGRA-	-CIRCLRRGK	A-CTYSI	N-
Phy000FWHP ASPFU					MAS-	TRDSHSYACDE	RLRKSR-	-CSK-ERPT	C-AQCKQLNK	EC-NYSP-H	-ITR-SP-L-
Phy000FX8T EMENIG	SMSEQ-R	- PSEPS - TPGS	KIPIPRVSQLR-	A-Y		GNRRVKRACIEC	R-KO-KTK-	C-N-GQTP-	-CSRCIGLGM	D - CVYMDG -	KREVT-
Phy000FXV0 EMENI			MDASVOPGHHLV	RA	- PAPGAEP -	SHRTAEOACREC	R-RR-KSKO	D-RAPV-	- CRLCSKY KR	0-CTYEP	-P-R-TP-L-
Phy000FYYQ EMENI MFIA	E/	A-A-NAS-KDL	ESGPRNPKSV	-T	GRTSTG-I	RRGVVPRACSSC		N-GEKP-	- CEACRWYKRPE	E-CTYPE-Y	(-ER
Phy000FZYA EMENI	PNDI	R-A-LIT-GAS	GGIGAACAHQLH	LA	- LTNS - AV -	NELVADLNS	-YA-DNKL	AGG-GING-	- CHYAASKGSTR	L-AEYNI-S	- VNP - RN - G-
Phy000G015 EMENI				M-	SSRGPT-I	RPRRL - AAC	HSOKTK-	-CSG-KR-P	C-DGCLSSGE	EC-QEPAK	E-RK-IR-V-
Phy000G06G EMENI GOYEAL	QVSPILRORST	L-A-VIG-GDL	LQNGHSA SE	AY	-GSVFRSR-I	HILRKTFACDEC	-RE-KIRC	S-GDEN-	- CLNCLRD AK	A-CRYSS-F	-SHQ-LS-K-
Phy000G0IY EMENI				M-	SKNMGS-	TRDSHSYAC			K0LDK	EC-KYSP-	-ITR-SP-L-
Phy000G1Z6 EMENI MFHT	E	G-P-AAS-APA	RAGRE		RQSSIG-	RRVTTLRACTSC	-HE-KING	D-GEKP-	- CEACRWYKKAD	L-CHYSD-F	-RPS
Phy000G28K EMENI			TDOPH	GAOPGIL	HASRLOGR -	KRKRVSVACRS	RA-R-KSR	-CN-GTOP-	C-SSCEDMDT	E-CRYDO-F	S-TRPRT-SV
Phy0006570 ASPNG					DISOSR-	RILRKTFACDE	K-RB-KIRO		- CNNCRRD AR	S-CRYSS-F	-SOR-LS-T-
Phy000G8Y0 ASPNG MYRS			S			KLSTTL					
Phy000G9MD ASPNG MASA	AD EDDSDSF	F-E-DHD-ASP	GHDLMMKDDAL -	GDGKP	DPLPMQ-I	KRRRVTRACDEC	R-RK-KIK-	CGK0P-	- CTHCTV YSY	E-CTYDO-F	SNR RRPA-
Phy000GA9V ASPNG				MNQT	HREGEPLA-I	KRRRIALACSAC	RA-R-KSR-	CD-GORPS	C-STCLSLTV	E-CVYEP-S	D-CSTNV-LV
Phy000GBJT ASPNG					MAS-	TRDSHSYACDE	RLRKSK-	-CSK-EKPT	C-AQCKQLNK	EC-NYSP-F	-VTR-SP-L-
Phy000GBKP ASPNGS	VPDGFPTRFNS	G-I-IP	STPTKLCNNIW-	LEMDS	GRSRGKS-I	RGKYT-KACEEC	R-B-RAK	- CD - GVKPS	C-SRCIQWNV	S-COYSG-T	E-DGRRP-A-
Phy000GEPP ASPNG	K	R			SEV-	OELRACVEC	R-RR-KSKO		-CSYCSRT KK	L-CREEH-A	A-P-R-TP-L-
Phy000H595 B0TFU		V-LST-E					AK-KIPO	H-IPYR-	- VDKAALAVINL	0-VIPDP-N	-PIR-NL-SY
Phy000H9RX BOTFU			-MEDK	DKDRD	RDTVNPKR-	RRLVT - NACTE	K-K-RAK	-CD-GET-P	C-ARCASOKV	E-CVYEV-F	V-ROSKE-N-
Phy000IALW COCIM MVP0	CAKSGANDO	V-A-SRT-APP	SOTPASADDNO-	AADDO	LAIPVO-	KRRRVTRACDEC	R-RK-KIK-	CGK0P-	- CTHCTV YSY	A-CTYDO-F	SNR RRPA-
Phv000IWR0_FUS0X		P-P0			A-I	PTVEEPTTCREC		D-GVPV-	-CSIC		R
Phy000IWT5 FUSOX			MFCT-E	PD	- SGTLKE - S	KPTRGLHACGNC	-RB-KIRC	D-GROP-	- CNOCESRDLOA	R-CVYTO-C	-IOR-VV-P-
Phy000J0D5 FUSOX	<mark>G</mark>	R			GSS-	SQLRACLNC	O-KR-KSRO	PVN-GPGP-	- CSYCARA GK	D-CSEVN-E	P-R-TR-L-
Phy000J1KA_FUSOX				<mark>E</mark>	LPF-SET-G	ROPRATSSCAE	R-RB-BVKO	CE-GQFP-	- CRQCVYYLVPH	L-CHYPA-F	R-KKR-HN-VS

statAI: Obtaining alignment statistics

While trimAl offers most of these options, this is a standalone program that will only give back some statistics of your alignment.

-sgc -sgt	Print gap score per column from input alignment. Print accumulated gap scores distribution from input alignment.
-ssc	Print similarity score per column from input alignment.
-sst	Print accumulated similarity scores distribution for input alignment.
-sfc	Print sum-of-pairs score per column for the selected alignment
-sft	Print accumulated sum-of-pairs scores distribution for the selected alignment
-sident -scolidentt	Print identity scores for sequences in the alignemnt. Print general descriptive statistics for column identity scores from input alignemnt.

All the exercises can be found in the folder named: trimal_tutorial. At the beginning of each exercise a tag will let you know in which subfolder you should be.

1.- **[example_readal]** Open the alignment file and check in which format it has been generated. Now use readal to (make sure each result is in a different file):

- Change the format of the current alignment to fasta format
- Change the format of the current alignment to nexus format
- Change the format of the current alignment to clustal format
- Use the -onlyseqs option

Open the different files and notice the differences between the alignment formats.

Tip: readal is run like this:

readal -in alignment_file -out trimmed_alignment_file -format FORMAT_NAME

You can check out all the formats supported by readal by typing: readal -h

2.- [example_trimal] Use trimAl to trim the alignments according to a gap threshold using the following parameters:

- A gap threshold of 0.1 (-gt 0.1)
- A gap threshold of 0.5 (-gt 0.5)
- A gap threshold of 0.9 (-gt 0.9)

Make sure that the output of your alignment is in phylip format. Now you can visualize each alignment either using a text editor or using seaview. Which of the previous commands deletes the largest amount of columns?

3.- **[example_trimal]** Now use the -gt 0.5 command but add a conservation score of different values: 30, 50 and 80 (-cons option). Again make sure that your output alignment is in phylip format. Which effect does it have on the trimmed alignment?

4.- **[example_trimal]** Now instead of using the gap threshold, we'll be using the similarity threshold (-st). Repeat the trimming of the original alignment using different similarity thresholds (0.1, 0.5 and 0.9). Again, how does the alignment trimming vary? Which approach is more aggressive? How can you make sure you don't loose all the alignment?

6.- [example_trimal] Now we are going to use the automated trimming methods. Trim your alignment using:

- Use the different automated trimming methods: -gappyout, -strict, -strictplus, -automated1
- Use the more radical methods to delete all the columns with gaps in your alignment: -nogaps

Of all the trimming strategies you've tried, which is the best one? Can you know?

7.- **[example_consistency]** You will see seven different alignments in there and a file called paths. Each alignment has been generated in a different way and we want to trim one of them based on the consistency score. Run this command:

trimal -compareset Phy007LWVO_COFCA.paths -forceselect Phy007LWVO_COFCA.alg.metalig -out Phy007LWVO_COFCA.alg.clean -phylip -ct 0.1667

This command will trim an alignment based on a set of alternative alignments. The -ct score will trim out columns that are inconsistent in the dataset. Which advantage do you think this kind of trimming has over the others we've seen?

8.- **[example_backtranslate]** Backtranslate protein alignment into CDS. In order to do this you need to use the -backtrans option:

trimal -in Phy007LWVO_COFCA.alg -out example.cds -backtrans Phy007LWVO_COFCA.cds

You can join this command with your preferred trimming methods.

9.- Trim all the alignments found in a folder. In this case you'll have to use a bit of bash programming:

- Create a new folder called trimmed_alignments
- Now move to the folder where you have all your alignments
- for fileName in \$(ls *); do trimal -in \$fileName -out ../trimmed_alignments/\$fileName -gt 0.1;done

(This will be needed in the main exercises after the explanations)







Concatenation & Partition Files

Rosa Fernández, Lisa Pokorny & Marina Marcet-Houben

Concatenation vs. Coalescence in Phylogenomics



Fig. 1. Liu et al. 2015. Ann. N.Y. Acad. Sci. 0:1–18.

Concatenation vs. Coalescence in Phylogenomics



Fig. 1. Liu et al. 2015. Ann. N.Y. Acad. Sci. 0:1–18.

Concatenation vs. Coalescence in Phylogenomics



Fig. 1. Liu et al. 2015. Ann. N.Y. Acad. Sci. 0:1–18.



Molecular Phylogenetics and Evolution

Volume 91, October 2015, Pages 98–122



Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms *

Mark P. Simmons^{a,} 📥 🔤, John Gatesy^b

ELSEVIER

^a Department of Biology, Colorado State University, Fort Collins, CO 80523, USA ^b Department of Biology, University of California, Riverside, CA 92521, USA

Received 12 February 2015, Revised 1 May 2015, Accepted 14 May 2015, Available online 19 May 2015

Syst. Biol. 63(6):919-932, 2014 © The Author(s) 2014. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved. For Permissions, please email: journals.permissions@oup.com DOI:10.1093/sysbio/sysu053 Advance Access publication July 30, 2014

Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies

Zhenxiang XI^1 , Liang Liu², Joshua S. Rest³, and Charles C. Davis^{1,*}

¹Department of Organismic and Evolutionary Biology, Haroard University, Cambridge, MA 02138, USA; ²Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA; and ³Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA; *Correspondence to be sent to: Department of Organismic and Evolutionary Biology, Haroard University, Cambridge, MA 02138, USA; E-mail: cdavis@oeb.haroard.edu.

> Received 7 January 2014; reviews returned 2 April 2014; accepted 24 July 2014 Associate Editor: Erika Edwards

Syst. Biol. 56(1):17-24, 2007 Copyright © Society of Systematic Biologists ISSN: 1063-5157 print / 1076-836X online DOI: 10.1080/10635150601146041

Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence

Laura Salter Kubatko^1 and James H. Degnan^2

¹Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA; E-mail: lkubalko@stat.ohio-state.edu
²Department of Biostatistics, Harvard School of Public Health, Building 2, 4th Floor, 655 Huntington Avenue, Boston, Massachusetts 02115, USA

Although multiple gene sequences are becoming increasingly available for molecular phylogenetic inference, the analysis of such data has largely relied on inference methods designed for single genes. One of the common approaches to analyzing data from multiple genes is concatenation of the individual gene data to form a single supergene to which traditional phylogenetic inference procedures—e.g., maximum parsimony (MP) or maximum likelihood (ML)—are applied. Recent empirical studies have demonstrated that concatenation of sequences from multiple genes prior to phylogenetic analysis often results

PLOS CURRENTS

Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions

March 9, 2015 · Systematics

Citation

Tonini J, Moore A, Stern D, Shcheglovitova M, Ortí G. Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. PLOS Currents Tree of Life. 2015 Mar 9. Edition 1. doi: 10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be.

Molecular Phylogenetics and Evolution

Syst. Biol. 56(1):17–24, 2007 Copyright © Society of Systematic Biologists ISSN: 1063-5157 print / 1076-836X online DOI: 10.1080/10635150601146041



Coalescence vs. of first principles app

Mark P. Simmons^{a,} ^a, ^a, Jo

Department of Biology, Univ

Received 12 February 2015, F

Syst. Biol. 63(6):919–932, 2014 @ The Author(s) 2014. Published by Oxford Univ For Permissions, please email: journals.permissio DOI:10.1093/sysbio/syu055 Advance Access publication July 30, 2014

LAB: TECHNIQUES FOR GENERATING PHYLOGENOMIC DATA MATRICES Toni / Marina / Rosa, 30 Jan

ed Data under Coalescence

NAN²

versity, Columbus, Ohio 43210, USA;

venue, Boston, Massachusetts 02115, USA

hylogenetic inference, the analysis common approaches to analyzing ergene to which traditional phylofL)—are applied. Recent empirical ohvlogenetic analysis often results

ethods Exhibit uracy under a

Coalescent versus Concatenation Methods and the Placement of *Amborella* as Sister to Water Lilies

Zhenxiang Xi¹, Liang Liu², Joshua S. Rest³, and Charles C. Davis^{1,*}

¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; ²Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA; and ³Department of Ecology and Evolution, Stony Brook University, Story Brook, NY 11794, USA; *Correspondence to be sent to: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; E-mail: cdavis@oeb.harvard.edu.

> Received 7 January 2014; reviews returned 2 April 2014; accepted 24 July 2014 Associate Editor: Erika Edwards

Range of Simulated Conditions

March 9, 2015 · Systematics

Citation

Tonini J, Moore A, Stern D, Shcheglovitova M, Ortí G. Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions, PLOS Currents Tree of Life. 2015 Mar 9. Edition 1. doi: 10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be.

Molecular Phylogenetics and Evolution

Syst. Biol. 56(1):17–24, 2007 Copyright © Society of Systematic Biologists ISSN: 1063-5157 print / 1076-836X online DOI: 10.1080/10635150601146041



Coalescence vs. of first principles app

Mark P. Simmons^{a, A}. ^{Ma}, Joi ^a Department of Biology, Color ^b Department of Biology, Unive

Syst. Biol. 63(6):919-932, 2014 ⊕ The Author(s) 2014. Published by Oxford Univ Yor Permissions, please email: journals.permissio DOI:10.1093/sysbio/syu055 Advance Access publication July 30. 2014 LAB: TECHNIQUES FOR GENERATING PHYLOGENOMIC DATA MATRICES Toni / Marina / Rosa, 30 Jan

Coalescent versus Concatenation Mathada and the Placement of Authonalia

ZHENX

¹Department of Organismic an Institute of Bioinformatics, Univer-

*Correspondence to be sent to

Species-tree estimation Laura Kubatko, 31 Jan ed Data under Coalescence

NAN

versity, Columbus, Ohio 43210, USA;

venue, Boston, Massachusetts 02115, USA

nylogenetic inference, the analysis common approaches to analyzing ergene to which traditional phylo-LL—are applied. Recent empirical shylogenetic analysis often results

ethods Exhibit uracy under a

ies Tree Methods



Phyloinformatic utility

Phyutility (fyoo-til-i-te) is a command line program that performs simple analyses or modifications on both trees and data matrices. Makes use of JADE (PEBLS) and JEBL libraries. Please see the NEWS page for info concerning updates, etc.

See NEWS and download the new release for amino acid acceptance in the concatenate and cleaning functions (use the -aa argument in -clean and -concat functions).

Please use this citation when using Phyutility Smith, S. A. and Dunn, C. W. (2008) Phyutility: a phyloinformatics tool for trees, alignments, and molecular data. Bioinformatics. 24: 715-716

Currently it performs the following functions (to suggest another feature, submit an Issue and use the label Type-Enhancement) :

Trees

- rerooting
- pruning
- type conversion
- consensus
- leaf stability
- lineage movement
- tree support

Data Matrices

- concatenate alignments
- genbank parsing
- trimming alignments
- search NCBI
- fetch NCBI



a) Concatenate the ortholog genes:

phyutility -concat -in INPUT_FILES -out OUTPUT_FILE

The output file will be in **<u>nexus format</u>**. It will include a block showing where each gene starts and ends.



a) Concatenate the ortholog genes:

phyutility -concat -in INPUT_FILES -out OUTPUT_FILE

The output file will be in <u>nexus format</u>. It will include a block showing where each gene starts and ends.

b) Create a partition file (command line with VIM).

Tutorial: Concatenation with Phyutility

1) Phyutility is a command line program written in Java. We'll concatenate the same genes that were aligned with Pasta and trimmed with trimAl.

If you had problems, you can use the pre-aligned genes in the folder '~/workshop_materials/orthologs_concatenation/*.aligned.fa'.

- 2) Open the terminal. You'll find the shortcut in the AMI Desktop.
- 3) Write the following commands:

phyutility -concat -in ~/workshop_materials/orthologs_concatenation/*.aligned.fa -out workshop_materials/orthologs_concatenation/concatenated_matrix.nexus

- 4) **concatenated_matrix.nexus** is your concatenated matrix in nexus format.
- 5) Write: vim concatenated_matrix.nexus and inspect the file (or nano concatenated_matrix.nexus).

How many taxa and amino acids does our matrix have?

For most phylogenetics/phylogenomics programs, you'll want to prepare a partition file. There are different ways of partitioning. Let's first partition **by gene**.

Let's do a file with partitions by gene by using our concatenated matrix (nexus format), as it contains a block indicating the starting and ending position of each gene.

1) Open the concatenated matrix with vim:

vim concatenated_matrix.nexus

- 2) Select with your mouse the block of partitions. Copy it (right click -> copy).
- 3) Create a new, empty file called **partitions.txt**

touch partitions.txt

4) Open the file:

vim partitions.txt (you can have a look at the Vim cheat sheet here)

- 5) Click 'i' to be able to edit the file. You'll see that you're now in INSERT mode.
- 6) Paste the block that you copied before (right click -> paste).
- 7) Remove the extension of the files:

:%s/.aligned.fa_//g

- 8) Now the names of all the genes are in the same line. Let's put them in different lines:
- :%s/ O/\r/g

9) Delete everything but the gene number:

:%s/G.*g/g/g

10) Insert the '=' symbol after the name of each gene:

:%s/ / = /g

11) Add the evolutionary model that you want to apply to each partition (eg., LG4X, WAG, etc.). In the new versions of RAxML and in ExaML, there's the option of automatic selection (AUTO).

:%s!^!AUTO, !

12) Save the file. Just type: :w

To quit vim, just type :q!

12) CONGRATS!! Your partition file is ready!!

Check the partition file. It should look like this:

You can go to the beginning of the file by typing 'gg', and to the end of the file with 'shift + gg'.

R Rc	sa					
D	Applica	ations	Places	System	۲	()
2						
File	Edit	View	Search	Termina	al Help	
AUTO, AUTO, AUTO,	gene1= gene2= gene3=	1-489 490-658 659-120	3 59			
AUTO, AUTO, AUTO,	gene4= gene5= gene6=	1270-18 1815-20 2015-21	314 914 186			
AUTO, AUTO, AUTO,	gene7= gene8= gene9=	2187-24 2424-26 2625-28	423 524 347			
AUTO, AUTO, AUTO,	gene10 gene11 gene12	=3453-4	4329 4104			
AUTO, AUTO,	gene15 gene15	=5008-	5543 5796			
AUTO, AUTO, AUTO,	gene10 gene17 gene18 gene19	=5967-0 =6513-0	5512 5722 5950			
AUTO, AUTO, AUTO,	gene20 gene21 gene22	=6951- =7560-8 =8848-9	7559 3847 9551			
AUTO, AUTO, AUTO,	gene23 gene24 gene25	=9552-9 =9753-9 =9911-	9752 9910 10521			
AUTO, AUTO, AUTO,	gene26 gene27 gene28	=10522 =10742 =10913	-10741 -10912 -11157			
AUTO, AUTO, AUTO,	gene29 gene30 gene31	=11158 =11346 =11914	-11345 -11913 -12074			
AUTO,	gene32 gene33	=12075	-12266 -12420			

Alternatively to partitioning by gene, we can also use some software to find best-fitting partition schemes. Let's try to find them with <u>PartitionFinder2</u>.

What PartitionFinder2 is for

PartitionFinder2 is a program for selecting best-fit partitioning schemes and models of evolution for nucleotide, amino acid, and morphology alignments. The user provides an alignment, and optionally some pre-defined data blocks (e.g. 9 data blocks defining the 1st, 2nd and 3rd codon positions of 3 protein-coding genes, see Figure 1). The program then finds the best partitioning scheme for this dataset, at the same time as selecting best-fit models for each subset of sites/columns. Here are a few things you can do with the program:

- 1. Find the best-fit partitioning scheme nucleotide, amino acid, or morphology datasets
- 2. Compare any number of user-defined partitioning schemes
- 3. Find best-fit models of evolution for each subset in any partitioned dataset (much like you might do with ModelTest or ProtTest).

The idea is that finding best-fit partitioning schemes and models of evolution will improve any downstream analyses of your data, like estimating phylogenetic trees or molecular dates. All of those kinds of analyses assume that your model of evolution is correct, and PartitionFinder2 helps make the model as good as it can be.

Let's copy the folder from the workshop webpage to your instance in the AMI.

1) Download the program from <u>this link</u>. We're doing this because it's not in the instance, so you can delete it from your computer after the practice. Also, it's a good exercise to learn how to use the 'scp' option to copy files or folders between a local host and a remote host.

Let's copy the folder from the workshop webpage to your instance in the AMI.

1) Download the program from <u>this link</u>. We're doing this because it's not in the instance, so you can delete it from your computer after the practice. Also, it's a good exercise to learn how to use the 'scp' option to copy files or folders between a local host and a remote host.

2) Open the terminal in your computer. Write the following:

scp -r path_to_folder phylogenomics@your_public_DNS:~/worskhop_materials

It will ask for the password: evomics2017

Let's copy the folder from the workshop webpage to your instance in the AMI.

1) Download the program from <u>this link</u>. We're doing this because it's not in the instance, so you can delete it from your computer after the practice. Also, it's a good exercise to learn how to use the 'scp' option to copy files or folders between a local host and a remote host.

2) Open the terminal in your computer. Write the following:

scp -r path_to_folder phylogenomics@your_public_DNS:~/worskhop_materials

It will ask for the password: evomics2017

3) Go to your AMI. Open the 'workshop_material' folder. Double-click the zip file and extract all the folders (or use tar -xzvf partitionfinder-2.1.1.tar.gz from your terminal).
4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

5) Go to the folder examples -> aminoacids. Open the **.cfg** file and have a look at it. There's information about each parameter in the documentation.

4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

5) Go to the folder examples -> aminoacids. Open the .cfg file and have a look at it. There's information about each parameter in the documentation.

6) Let's run an example. In your AMI, open a terminal. Go to **'workshop_materials ->** partitionfinder_2.1.1"

4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

5) Go to the folder examples -> aminoacids. Open the **.cfg** file and have a look at it. There's information about each parameter in the documentation.

6) Let's run an example. In your AMI, open a terminal. Go to **'workshop_materials ->** partitionfinder_2.1.1"

7) Type the following:

python PartitionFinderProtein.py examples/aminoacid/

4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

5) Go to the folder examples -> aminoacids. Open the **.cfg** file and have a look at it. There's information about each parameter in the documentation.

6) Let's run an example. In your AMI, open a terminal. Go to 'workshop_materials -> partitionfinder_2.1.1"

7) Type the following:

python PartitionFinderProtein.py examples/aminoacid/

8) Once its finished, go to 'examples -> aminoacid' and explore the output: **log.txt** and a folder called **analysis.**

4) Navigate to the extracted folder. You'll see several python scripts. You'll need to chose the correct one depending on the type of data that you have (morphology, nucleotides or amino acids).

5) Go to the folder examples -> aminoacids. Open the **.cfg** file and have a look at it. There's information about each parameter in the documentation.

6) Let's run an example. In your AMI, open a terminal. Go to 'workshop_materials -> partitionfinder_2.1.1"

7) Type the following:

python PartitionFinderProtein.py examples/aminoacid/

8) Once its finished, go to 'examples -> aminoacid' and explore the output: **log.txt** and a folder called **analysis.**

9) In the folder analyses, you have the file **best_scheme.txt.** Have a look at it.

Putting it all together...

Exercise: alignment, trimming, concatenating, creating partition files

In the folder 'workshop_materials -> ortholog_concatenation' there are ~ 100 1:1 orthologs. They're called **OG[NUMBER].fa**

Following the same steps that you've learned in the labs today

- align them with PASTA
- trim them with trimAL
- create a concatenated matrix
- create a partition-by-gene file, and
- create a best-fit partition file with PartitionFinder2 (you'll need to change to phylip format with SeaView)