User guide

LS^{x} v1.1

13.11.2017

A script in R to run the LS³ and LS⁴ phylogenetic data subsampling algorithms for reducing lineage rate heterogeneity

Carlos J. Rivera-Rivera[1,2]
Juan I. Montoya-Burgos[1,2]

[1]*Department of Genetics and Evolution, University of Geneva, Switzerland*
[2]*Institute of Genetics and Genomics in Geneva (iGE3)*

Table of Contents

1 Quick Start

1.1 Download and installation

Download the R script `LSx_v1.1.R` and the LS^X input file `LSx_input_file.txt` from https://genev.unige.ch/research/laboratory/Juan-Montoya (under the “MORE” tab).

DONE! (LS^X runs directly through the `Rscript` or `Rscript.exe` scripting front-end, see point 1.3)

Check dependencies in point 3.

1.2 Input files

- **Gene sequence alignments**, each as an independent file in PHYLIP interleave format. Taxa names of less than 10 characters are identical across gene datasets, unique within a gene dataset, and no name is a subset of another name (format details in section 4.1).
- The **list of gene alignment files** to be analyzed, with (optionally) the model of sequence evolution; in raw text, with comma-delimited columns if needed (format details in section 4.2).
- **Guide tree** in Newick format, rooted, with a polytomy at the base of the lineages of interest (format details in section 4.3).

- **Lineage-taxon file** in raw text format, with comma-delimited columns (format details in section 4.4).
- **LS^X input file** with the values for your current run (details in section 4.5).
- **PAML control files**. These are baseml.ctl or codeml.ctl (aaml.ctl in the LS^X example files).

1.3 Running

Linux/Mac command:

```
Rscript LSx_v1.1.R LSx_input_file.txt
```

Windows command:

```
Rscript.exe LSx_v1.1.R LSx_input_file.txt
```

2 Short Introduction

2.1 What is LS^X?

LS^X is a script in R that runs the LS³ and LS⁴ algorithms of data subsampling for multi-gene phylogenetic inference. Both of these algorithms do a gene-by-gene inspection of the heterogeneity of evolutionary rates among user-defined lineages of interest (LOI). Then, using criteria that differ in both algorithms (please refer to the publications or section 2.2 below for more details), they try to find a subsample of sequences that evolve at a homogeneous rate across all LOIs. If this subset is found, an alignment of the gene is produced with only the sequences that evolve homogeneously. At the same time, a table is also produced showing which sequences were "flagged" (the sequences that were removed), and which sequences were kept. If a subset of sequences that evolve at a homogeneous rate is ***not*** found, the gene is flagged entirely.

Originally, a set of bash scripts were published to run LS³ on Linux computers (Rivera-Rivera and Montoya-Burgos 2016), but they were not very accessible nor user-friendly, and lacked some important options. Also, in those scripts, it was not possible to run LS⁴, as this algorithm was developed later. We hope to have changed this by completely re-programming everything in R, including LS⁴.

LS^X relies on PAML (Yang 2007) for optimizing the likelihood values, so it's very useful to know at least a little about PAML in order to understand the power and the limitations associated to its maximum likelihood calculations.

2.2 Is LS^X for me?

Both algorithms included in LS^X are for cases in which you are certain of the monophyly of 3 or more lineages, but you suspect that the relationships among them may be affected by artifacts due to lineage evolutionary rate heterogeneity, including long branch attraction

(LBA). Your LOIs should comprise all of the ingroup species in the dataset. If you have some stray ingroup taxon/taxa that do/does not fall within any of your LOIs and whose position among your LOIs is not consequential, you can put it aside for the LS^X analyses. If it *is* consequential, then, include it as a LOI on its own.

That said, while the amount of LOIs allowed is unlimited in the code, we recommend not to evaluate 5 or more LOIs at a time. As explained below, the input tree for the LS^X analyses must have a polytomy at the base all of your LOIs, in order not to impose a particular branching arrangement when calculating the likelihood of the different models. Thus, the more LOIs declared, the larger the polytomy will be, and we have observed that this can result in strange allocations of substitution information.

If you have 5 or more monophyletic lineages whose interrelationships may be affected by lineage rate heterogeneity, you can still run LS^X! Simply, analyze subtrees as 3- or 4-LOI cases, and then consolidate the information.

2.3 LS³ or LS⁴?

Briefly, LS³ will only consider long branches to be disrupting of lineage rate heterogeneity, whereas LS⁴ will also consider short branches. We have noticed that LS³ can be more stringent, and can flag more genes than LS⁴. We use both.

3 Dependencies

LS^X needs the following to be present/installed in your computer:

- R (make sure also that you also have the `Rscript/Rscript.exe` scripting front-end)
- PAML (you must know where the `baseml/baseml.exe` and/or `codeml/codeml.exe` binaries are)
- R's *ape* package (v. 3.X)
- R's *adephylo* package
- R's *parallel* package, if you want to run it in parallel (parallel analysis is functional but under development)

4 Input data

4.1 Sequence alignments

LS^X reads sequence data in PHYLIP interleave format. Both LS³ and LS⁴ algorithms approach the sequence subsampling in a gene-by-gene manner, so independent alignment files are expected for each gene.

Important things about input alignments:

- They must be aligned (i.e., all sequences must be of the same length).
- The taxa/OTU names must be unique within an alignment, and match across all gene alignment files.
- Avoid at all costs taxa/OTU names that are subsets of others. For example, "**Celegans**" and "**Celegans**1" will give trouble, but "Celegans1" and "Celegans2" won't.
- We **highly** recommend that the species names are of 10 characters maximum. This avoids any problem of name changing due to truncation by some software.

4.2 Alignment file table with (optional) PAML models of sequence evolution

In order to tell LS^X which gene alignments to analyze, you need to give it the list of filenames. This is a plain text file, and if more than one column is added (see Example 2 and 3 below), columns are separated by commas. In it, you list the file names, and also if you want to give some information to PAML on the model of sequence of evolution to be used.

Example 1: an alignment input file for the analysis of nucleotide or amino acid data, for which you want to use the default sequence evolutionary model parameters. The default for nucleotides is GTR+G with 4 rate categories, and the default for amino acids is the rate matrix given by the `DefaultAAMatrix` value (user-defined, see below), also with the gamma model of site heterogeneity with 4 rate categories.

```
gene1.phy
gene2.phy
gene3.phy
gene4.phy
gene5.phy
```

Example 2: an alignment input file input for the analysis of nucleotide data, for which you want to specify the sequence evolutionary model to PAML:

```
gene1.phy,1,,,
gene2.phy,5,G,8
gene3.phy,7,G,10
gene4.phy,7,G,10
gene5.phy,7,G,10
```

The first column contains the alignment's file name, the second column has the code for the model of sequence evolution as used by PAML (see Table 1 below), the third column contains a request of the gamma distribution model of site rate heterogeneity (leave blank if the gamma distribution is not needed), and the fourth column, how many gamma site rate categories you want (also leave blank if the gamma distribution is not needed). Notice

that when the gamma model of site heterogeneity is not needed, the 3rd and 4th columns are left blank, but the commas must be there.

Table 1. Codes for models of nucleotide evolution used in PAML (more info in PAML's manual).

Sequence Evolution Model	PAML's Code Number
JC69 (a.k.a. Jukes-Cantor)	0
K80 (a.k.a. Kimura 2-parameter, K2P)	1
F81	2
F84	3
HKY85 (a.k.a. HKY)	4
T92	5
TN93 (a.k.a. Tamura-Nei 93)	6
REV (a.k.a. GTR)	7
UNREST	8
REVu	9
UNRESTu	10

In the example shown above, gene1 uses the Jukes-Cantor model without employing the gamma distribution to estimate site rate heterogeneity, gene2 uses the T92 model with the gamma distribution, with 8 site rate categories, and the last three genes use the GTR model with the gamma distribution, and 10 rate categories.

Example 3: an alignment input file input for the analysis of amino acid data, for which you want to specify the sequence evolutionary model to PAML:

```
gene1.phy,wag.dat,G,8
gene2.phy,jones.dat,G,4
gene3.phy,mtmam.dat,,,
gene4.phy,lg.dat,G,8
gene5.phy,lg.dat,G,8
```

PAML doesn't have a code for the amino acid model, but instead has files with the matrices that are used. The name of the file of the matrix you want to apply goes in the second column (these matrices are included in the downloaded package of PAML). The third and fourth columns are, like before, whether you would like to use the gamma distribution to model site rate heterogeneity, and if so, how many categories you want to use. Hence, gene1 uses the WAG matrix with the gamma distribution, and 8 site rate categories, gene2 uses the JTT matrix, with the gamma distribution and 4 rate categories, gene 3 uses the mtMAM matrix, without modeling site rate heterogeneity, and the last two use the LG matrix with the gamma distribution and 8 site rate categories.

4.3 Guide tree for likelihood estimation

The guide tree is in Newick (parenthesis) format, in which the monophylies of the ingroup, the outgroup, and all of the LOIs are present, and it must be rooted. The other relationships need not be resolved. However, we recommend to add as much topological information as possible, especially the relationships among the outgroups. However, the essential point about this tree is that the relationship among your LOIs is a polytomy (see schematic example in Figure 1).

The names of all taxa/OTUs have to be identical to the names they have in the alignments, and this tree should contain all possible taxa/OTU that will be analyzed. In other words, all of the taxa/OTUs in the alignments have to be in the tree (although not all of the taxa/OTUs in the tree have to be in all of the alignments).

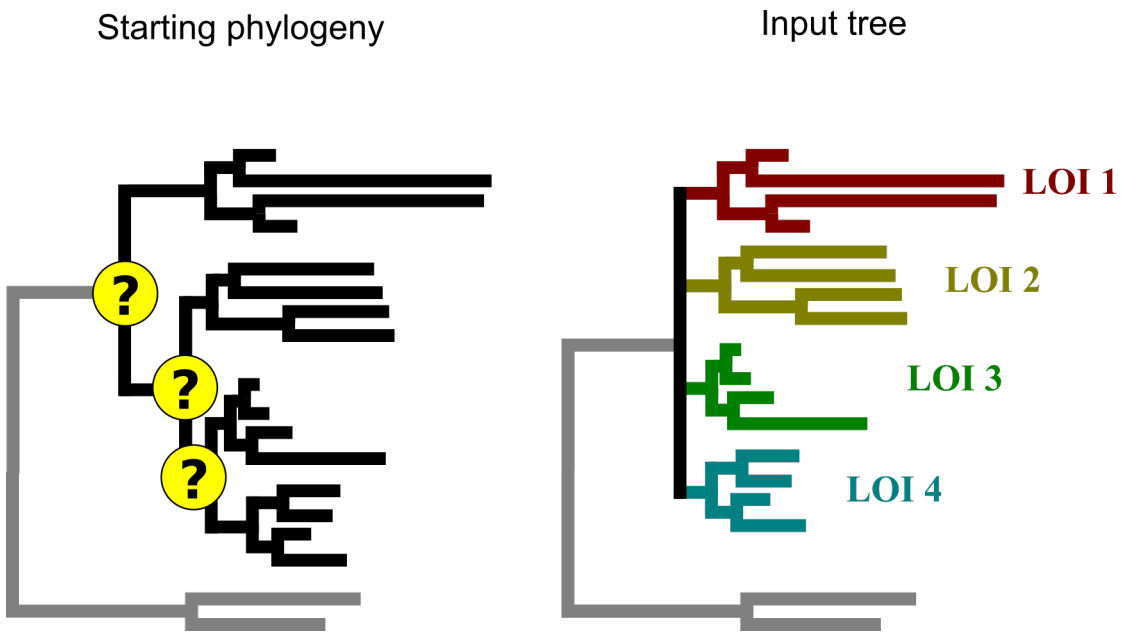

Figure 1. A schematic representation of the input tree needed for LS^X analyses. In the left, a phylogeny where the placing of some lineages may be the result of lineage rate heterogeneity (nodes marked with a "?"). In the right, the input tree given to LS^X to reduce lineage rate heterogeneity. Four LOIs are declared, and the relationship among them is a polytomy.

4.4 Lineage-taxon file

This file is a table that states to which LOI each taxon/OTU belongs. It is in raw text, with columns delimited with commas (as is the gene input table). On the first field is the LOI to which a taxon/OTU belongs, and the second field is the name of that taxon as it appears in the alignments and input tree.

Example:

```
Clade1,Species1
Clade1,Species2
Clade2,Species3
```

```
Clade2,Species4
Clade3,Species5
Clade3,Species6
Clade3,Species7
Clade4,Species8
Ogs,Species9
Ogs,Species10
```

While the formatting of the species name is not constrained (except for the points mentioned in 4.1), the name of the LOI must be as shown in the example, i.e. Clade1, Clade2, Clade3, Clade4, etc. Also, the outgroups must be identified as “Ogs”.

4.5 LS^X input file

This is the core file of instructions for LS^X, and contains some variable declarations that will pass directly into R. It is important that each variable entry is in “double quotes”, as shown in the example below.

Example:

```
# LS3/LS4 R input file
# Remember that all input is to the LEFT
# of the equal sign (=) and MUST be in
# "double quotes"

listFile = "Gene_Input_List.txt"
guideTree = "Input_Tree.nwk"
CladeSpeciesFile = "Lineage_taxon_File.txt"
GenericPamlCtrlFile = "aaml.ctl"
PamlExecutablePath = "/home/user/bin/codeml"
dataType = "NUC"
defaultAAMatrix = "wag.dat"
Flavor = "LS4"
pThresh = "0.05"
minTaxa = "1"
numCores = "3"
```

Details:

listFile	The filename of the text document in which LS^X will find the list of genes to be analyzed, with the (optional) model parameters. This is the document explained in point 4.2.
guideTree	The filename of the document in which the guide tree (from point 4.3) is.

CladeSpeciesFile

The filename of the lineage-taxon file (from point 4.4).

GenericPamlCtrlFile

The filename of the generic PAML control file. In the original PAML folders, these are "baseml.ctl" for nucleotides, and "codeml.ctl" for amino acids. In the examples provided with LS^{X}, these are "baseml.ctl" and "aaml.ctl", for nucleotides or amino acids, respectively.

PamlExecutablePath

This is the path to the PAML executable. For an analysis of nucleotide data this is `baseml` or `baseml.exe` (depending if you are on a UNIX or a Windows computer, respectively), and for amino acid data, `codeml` or `codeml.exe`. This must be the absolute path, so, on a UNIX computer it will look something like the example above (starting from your root folder, i.e. "/"). In Windows computers, paths use backslashes ("\"), and for giving file addresses into R, they must be doubled. For example, a path would look more like: "`C:\\User Files\\Program Files\\paml4.7\\baseml.exe`"

DataType

There are two options here: "NUC", if your alignments are in nucleotide or "AA" if they are in amino acid. Both datatypes cannot be mixed in the same LS^{X} run (but can be done separately).

DefaultAAMatrix

This value is important only if you are running the analysis with amino acid data and you decide to use a default amino acid substitution matrix (as in 4.2, *Example 1*). LS^{X} will look into this field in order to know the filename of the PAML amino acid matrix that you want to use as default. This variable is ignored (and you can comment it out) if you are doing an analysis of nucleotides, or of amino acids but giving the specific matrices for each gene.

Flavor

Here you choose whether you want to run the sequence subsampling algorithm following the LS^{3} ("LS3") or LS^{4} ("LS4") criterion.

Pthresh

LS^{X} uses a likelihood ratio test (LRT) for determining whether a current sequence subselection evolves at a homogeneous rate. The main value that determines whether the

	heterogeneous lineage rates and the homogeneous lineage rate models are equal or different is the p-value resulting from the LRT. We have always used a p=0.05, but other p-values can be used. Using p-values higher than 0.05 will result in more stringent enforcement of lineage rate homogeneity. We do not recommend using lower values.
`MinTaxa`	When a LOI reaches this amount of taxa, no more sequences will be removed from this lineage. LOIs that have *a priori* less sequences than this number will not be touched.
`NumCores`	If you want the analysis to be parallelized, enter here the amount of cores to be used. If this entry is "1", or commented out, LS^X will run sequentially, on a single core. ***Note: This function is still giving problems with certain computers***.

4.6 PAML control file

This is the "baseml.ctl" or "codeml.ctl" files of PAML. In our examples, we provide them as "baseml.ctl" and "aaml.ctl".

5 Running

LS^X launches from the command line, both in Linux/Mac and in Windows, and takes as only argument the name of the LS^X input file.

Command for Linux/Mac:

```
Rscript LSx_v1.1.R LSx_input_file.txt
```

Command for Windows:

```
Rscript.exe LSx_v1.1.R LSx_input_file.txt
```

Note: These commands are if your Rscript program is in your PATH. If not, just give the full address to the executable.

6 Output data

LS^X will produce several files and folders. For a given gene "gene1", it will produce:

- A folder called "gene1_LS3" or "gene1_LS4" (depending on which variation of the algorithm was used), in which you will find the alignments for all data subsamples

taken for that gene (each subsample is prepended with “ss”, so, for example, the alignment of subsample 22 will be “gene1_ss22.phy”). Also, for each subsample, you will find the backbone tree used for branch length optimization (e.g. “gene1_ss22_backbone.nwk”), the same tree with optimized branch lengths (e.g. “gene1_ss22_branchopt.nwk”), the tree with branch lengths optimized assuming a model of homogeneous lineage rates (e.g. “gene1_ss22_SR.nwk”), and the tree with branch lengths optimized assuming heterogeneous lineage rates (e.g. “gene1_ss22_MR.nwk”). Also, all of the PAML output files for these analyses will be there.

- An alignment called “gene1_homRatesLS3.phy” or “gene1_homRatesLS3.phy” (again, depending on the type of analysis) that includes only the sequences that were found to evolve at a homogeneous rate. If no such sequence subset can be found for that gene, this file will not be present.
- A table in comma-separated file (this format can be read in OpenOffice, LibreOffice, Excel, R, …) called “gene1_LS3Out.csv” or “gene1_LS4Out.csv”, depending on the type of analysis. In this table you will find a summary of all the information of the sequence subsampling process for that gene and all subsampling step, including the rates and relative rate of all lineages, the p-value of the LRT, and the sequences removed.

In addition, LS^X will produce a table for the entire set of genes given in the input gene list (the one from point 4.2). It is a comma-separated file, and it shows for each gene, which taxa/OTUs were kept (“1”) and removed/flagged (“0”) in order to reach lineage rate homogeneity. Taxa/OTUs that were originally not present in a gene are shown as “NA”. Genes that were flagged entirely are evident in this table because all of their taxa/OTUs are shown as “0”.

Example (fragment):

	Species1	Species2	Species3	Species4	Species5	Species6
gene1.phy	1	0	0	1	0	1
gene2.phy	1	1	0	1	1	1
gene3.phy	1	1	1	1	1	1
gene4.phy	0	1	1	0	1	1
gene6.phy	0	0	0	0	0	0
gene7.phy	0	0	0	0	0	0
gene8.phy	0	0	0	0	0	0
gene9.phy	1	1	1	1	0	1
gene10.phy	1	NA	0	0	NA	1

The filename of this table starts with “LS3_presAbsTot” or “LS4_presAbsTot” (depending on the type of analysis), followed by a time stamp of when it was produced.

7 Limitations

With this reprogrammed version of LS³ and the first version of LS⁴ we have addressed many of the main shortcomings from the initial bash scripts, several of them thanks to the input of users. However, there are still some limitations that are inherent to the calculations run in this script, and these must be taken into account by the user when running an analysis.

- Gene alignments must be informative enough. The sequence-specific evolutionary rates cannot be calculated well in datasets containing little or no substitutions.
- In all alignments, all LOIs as well as outgroups must be represented by at least one sequence.

8 Citing LS^{X}

If you use LS^{X} for LS⁴, please cite:

Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS^{X} : Automated reduction of gene-specific lineage evolutionary rate heterogeneity for multi-gene phylogeny inference. BiorXiv (2017). DOI: 10.1101/220053

If you use it for LS³, please also cite:

Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS³: A Method for Improving Phylogenomic Inferences When Evolutionary Rates Are Heterogeneous among Taxa. *Mol. Biol. Evol.* 33, 1625–1634 (2016). DOI: 10.1093/molbev/msw043

Thank you!

9 References

Jombart, T., Balloux, F. & Dray, S. Adephylo: New Tools for Investigating the Phylogenetic Signal in Biological Traits. *Bioinformatics* 26, 1907–9 (2010). DOI: 10.1093/bioinformatics/btq292

Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290 (2004). DOI:10.1093/bioinformatics/btg412

Popescu, A. A., Huber, K. T. & Paradis, E. Ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* 28, 1536–1537 (2012). DOI: 10.1093/bioinformatics/bts184

Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS³: A Method for Improving Phylogenomic Inferences When Evolutionary Rates Are Heterogeneous among Taxa. *Mol. Biol. Evol.* 33, 1625–1634 (2016). DOI: 10.1093/molbev/msw043

Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS^{X} : Automated reduction of gene-specific lineage evolutionary rate heterogeneity for multi-gene phylogeny inference. BioRxiv (2017). DOI: 10.1101/220053

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–91 (2007). DOI: 10.1093/molbev/msm088. You can download it here.