2024 Historian Evaluation Procedure

Krishna Bhatt @ Holmes Lab, UC Berkeley 2024

Table of Contents

Table of Contents 1
Investigation Summary 2
Alignment Simulation with indel-Seq-Gen 2.0 3
Download and Installation 3
Tool Overview 3
Retrieving ‘Ancestral’ Sequences 4
Setting Simulation Parameters 6
Generating Trees Structures 7
Simulating Phylogenetic Trees 7
Historian & BAli-Phy 3 Evaluation 8
Reconstructing Ancestral Trees 8
Running Historian & BAli-Phy 8
Model Fitting for Parameters 9
Developing Representative Tree Metrics 9
Download and Installation 10
Tree Analysis and Comparison 11
Evaluating MCMC Mixing 11
Data Investigation 12
Aggregated References 13

Investigation Summary

This evaluation will focus on comparing Historian to another ancestral sequence reconstructor, BAli-Phy. Both are unique to other phylogenetic reconstructors in that they sum over all possible alignments rather than basing all further computation on the single most probable alignment. In addition, the two algorithms prioritize evolutionary accuracy rather than structural similarity in their reconstructed trees with Historian utilizing an adaptation of ProtPal and explicit evolutionary parameters.

In order to test the viability and accuracy of these two programs, they will be put through a sequence of tests. A number of biologically diverse modern sequences will be retrieved for use as an "ancestral sequence." Then based on realistic parameters, a future tree will be simulated with indel-Seq-Gen. Historian and BAli-Phy will reconstruct the sequences while being measured for wall-clock time. Finally, metrics such as a Generalized Robinson Foulds algorithm and "Align" will be utilized to compare the quality of results. MCMC mixing will also be evaluated with Mean Compute Time Statistics.

The entire evaluation will be centered around a mapping file which contains information about each "Ancestral Sequence," their indel and substitution rates, their reconstructed trees, as well as the performance metrics and wall-clock time for reconstruction. Keeping with this organization, the files should all primarily be converted to CSV, FASTA, or NEWICK file formats if possible.

I have tried to ensure that my evaluation plan is as specific as possible while being simple to follow, but if you have any questions regarding any part of the document, please reach out to me at krishbhatt2019@gmail.com.

Trello Board Task:

The Historian software includes code for reconstructing phylogenetic trees using joint MCMC sampling over trees and alignments. The nearest competitor is BAli-Phy. The goal of this (rather advanced) task is to evaluate how good Historian is at this, and compare it to BAli-Phy.

This will involve:

- reviewing the historian code base, including the (minimally documented) options for phylogenetic reconstruction
- writing code to simulate a bunch of alignments using a tool such as indel-seq-gen
- writing wrapper scripts to reconstruct the trees by calling Historian or BAli-Phy, measuring the amount of wall-clock time taken (MCMC is slow and speed/performance will be an important part of the evaluation)

- assessing the accuracy of the reconstructed trees, e.g. using the Robinson-Foulds metric
- assessing the ability to retrieve simulation parameters, as in this BAli-Phy paper
- assessing the mixing properties of the MCMC run using the methods outlined in this paper

This task is probably graduate-level, but is included here in case any undergrad or HS interns want to take a look!

Alignment Simulation with indel-Seq-Gen 2.0

Key Resources (indel-Seq-Gen):

- https://pubmed.ncbi.nlm.nih.gov/17158778/ **[Paper]**
- http://bioinfolab.unl.edu/~cstrope/iSG/#Software **[Download]**
- http://bioinfolab.unl.edu/~cstrope/iSG/iSGv2_manual.pdf **[Manual]**

Download and Installation

There are a couple of changes that have to be made to the indel-Seq-Gen v.2.1.03 such that it runs on a modern system (tested on Ubuntu Linux). Primarily, setup follows the steps described in the manual, however some files have to be modified for it to install successfully.

After downloading and unzipping required files from the homepage, make the following changes before running `./configure` and `make`:

1. In **main.cpp**, add the following line: `#include <unistd.h>`
2. In **nucmodels.cpp**, replace variable name `beta` with an arbitrary alternative such as `nbeta`
3. In **random.cpp**, replace variable name `next` with an arbitrary alternative such as `nnext`

After these steps, you can continue with the normal installation steps, and it should proceed as expected.

Tool Overview

4 Basic Environments:

1. Global Environment: Sets the default substitution parameters for simulation run
2. Partition Environment: Defines partition-specific simulation parameters
3. Subtree Environment: Defines lineage-specific simulation parameters
4. Motif Environment: Defines Site-specific simulation parameters

	Substitution Parameters
θ^m	Substitution Matrix
θ^f	Character Frequencies
θ^r	Site Rates
θ^i	Percent Invariable Sites

	Indel Parameters
λ^p	Probability of indel at the number of indels per substitution
λ^l	Length distribution
λ^m	Maximum length

To run indel-Seq-Gen:

```
indel-seq-gen -m <matrix> [-options] < guide_tree_file > outfile
```

NOTE: I created this section for personal notes but figured that it may be useful.

Retrieving 'Ancestral' Sequences

To act as 'ancestral sequences' during this simulation, **modern sequences** will be utilized to maximize the test group and the rigor of the test due to the limited availability of real diverse phylogenetic trees. A total of 12k sequences will be selected from NCBI's GenBank with the following sub-divisions:

| **Bacteria (4k Total Sequences**
| **Animals (4k Total Sequences)**
- Mammals (1k Sequences)
- Fish (1k Sequences)
- Birds (1k Sequences)
- Insects (1k Sequences)

| **Plants (4k Total Sequences)**
| **Viruses (4k Total Sequences)**
- RNA Viruses (2k Sequences)
- DNA Viruses (2k Sequences)

In order to make tests as realistic as possible, the reconstructed trees will be divided into 3 classes (Bacteria, Animal, and Virus) to represent the 3 major use-cases for Historian. As such the reconstructed trees will also be developed to match certain tree characteristics prominent in each of the respective classes (Tree Structure, Branch Lengths, Substitution Rates, Indel/Mutation Rates, etc.) Further distinction is also made in the Animal and Virus class for a more detailed evaluation.

Using BioPython's Entrez Module, extract 12k DNA sequences from GenBank following the subdivisions in the table above.

NOTE: Retrieving NCBI GenBank data may require an API key obtained from their website.

Rules for Extraction:

1. 12,000 sequences in total divided amongst the three classes
2. Batched and time-staggered approach to downloading (each class forms its own batch)
3. Sequences are 500 to 5000 base pairs in lengths

4. Diversity filter to only keep one sequence per genus

Each gene would be sorted into its own individual FASTA file, all within a separate folder per batch. Additional meta-data files should be generated per batch to outline details (gene length distribution, no. genes, gene taxonomic information, etc.) for easy mapping and reference. They should be organized in a manner similar to such:

```
seqs/
├── bacteria/
│   ├── info.csv
│   ├── seq_1.fasta
│   ├── seq_2.fasta
│   └── ...
├── plants/
│   ├── info.csv
│   ├── seq_1.fasta
│   ├── seq_2.fasta
│   └── ...
├── animals/
│   ├── mammals/
│   │   ├── info.csv
│   │   ├── seq_1.fasta
│   │   ├── seq_2.fasta
│   │   └── ...
│   ├── fish/
│   ├── birds/
│   └── insects/
├── viruses/
│   ├── dna/
│   │   ├── info.csv
│   │   ├── seq_1.fasta
│   │   ├── seq_2.fasta
│   │   └── ...
│   └── rna/
```

The specific genes within each class may be arbitrary or up to your judgment.

Setting Simulation Parameters

As the evaluation is trying to test realistic parameters across a wide variety of taxonomic groups, various parameters will be pseudo-manually set to ensure realisticness. This will include, indel/substitution rates, tree structure, branch lengths, etc. Further specification can be created

upon necessity. Some example data for the 'Animals' taxonomic group is shown below. Similarly, this process of determining parameters will have to be repeated for all taxonomic groups.

Taxo. Group	**Indel Rates**	**Substitution Rates**
Mammals	2.2×10^{-9} mut/bp/year	2.22×10^{-9} subs/site/year
Birds	4.6×10^{-9} mut/site/gen	0.683 subs/site/billion years
…	…	…
DNA Virus	1.8×10^{-8} mut/nt/rep	1.5×10^{-3} subs/site/year

Here are some good starting points for research:

1. Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A. 2002 Jan 22;99(2):803-8. doi: 10.1073/pnas.022629899. Epub 2002 Jan 15. PMID: 11792858; PMCID: PMC117386.

2. Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird. Genome Res. 2016 Sep;26(9):1211-8. doi: 10.1101/gr.204669.116. Epub 2016 Jul 13. PMID: 27412854; PMCID: PMC5052036.

3. Benoit Nabholz, Axel Künstner, Rui Wang, Erich D. Jarvis, Hans Ellegren, Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics, *Molecular Biology and Evolution*, Volume 28, Issue 8, August 2011, Pages 2197–2210, https://doi.org/10.1093/molbev/msr047

4. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008 Apr;9(4):267-76. doi: 10.1038/nrg2323. Epub 2008 Mar 4. PMID: 18319742.

5. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. Proc Natl Acad Sci U S A. 1999 Oct 26;96(22):12638-43. doi: 10.1073/pnas.96.22.12638. PMID: 10535975; PMCID: PMC23026.

6. T Heath Ogden, Michael S Rosenberg, Alignment and Topological Accuracy of the Direct Optimization approach via POY and Traditional Phylogenetics via ClustalW + PAUP*, *Systematic Biology*, Volume 56, Issue 2, April 2007, Pages 182–193, https://doi.org/10.1080/10635150701281102

When simulating the data, the rates should not be static, but they should be varied in one standard deviation off of the mean to introduce randomness to the data set. To save the unique variations and the associated 'Ancestral Sequence,' simulated parameters should be mapped to their corresponding sequences in a CSV, which includes the following information: **mapping to**

sequence, taxonomic group, indel rate, substitution rate, tree length, no. of leaves, and mapping to tree.

The depth of the tree can be 25 with a standard deviation of 7. The number of leaves can be 40 with a standard deviation of 25. These numbers are completely flexible and estimates, so they can be changed in the future due to testing scenarios. The tree generation will be discussed in the next section.

As for the indel model, we will be using the **GP120 model**, and for substitution, **Dayhoff PAM matrix**- the same models used for the first evaluation (Holmes, I. 2017).

For another possible test, we could also vary the substitution and indel models. As per current studies (Zardoya, R. 2021), HKY85 is a robust choice of a substitution model for its ability to both allow unequal nucleotide frequencies as well separate rates for transitions and transversion. TKF92 is an update to TKF91's restriction of single residue indels by using 'fragments' instead of residues. A similar functionality is also presented by the "Long-indel" model in this paper (Miklós I. et. al, 2004).

Generating Trees Structures

Similar to how the simulation parameters have to be set to each taxonomic group, the details for the tree structure have to be based on realistic constraints, the taxonomic structure has to be based in reality. For that, each taxonomic group's average birth and death rate have to be collected for the simulation. They should also be varied one standard deviation off of the mean.

Tree generation can be handled by the python library ETE in conjunction with DendroPy to create birth-death processes and coalescent models.

Simulating Phylogenetic Trees

Since indel-Seq-Gen doesn't support GPU-based parallelization, the runtimes should be separated into batches. The most intuitive partitioning of batches would be to run each taxonomic group on separate machines/CPUs/cores in parallel. However, these simulations aren't expected to take as much time as the reconstruction algorithms or the accuracy analysis, so the time should not be much of a concern.

That being said, if the computations are too extensive, a non-discrete Gillespie algorithm (GIL), can be used, as described in the manual.

All simulations should be run by programmatically looping through the CSV and generating corresponding trees and MSAs.

The output root sequences, sequence files, multiple alignments, and traced events should be stored together in one folder within a directory of similar structure to the sequence directory. Therefore, the previously created CSV file should be cross-compatible with this new directory.

Historian & BAli-Phy 3 Evaluation

Reconstructing Ancestral Trees

Key Resources (Historian):

- https://doi.org/10.1093/bioinformatics/btw791 **[Paper]**
- https://github.com/evoldoers/historian **[Github]**

Key Resources (BAli-Phy):

- https://doi.org/10.1093/bioinformatics/btab129 **[Paper]**
- https://www.bali-phy.org/ **[Download]**
- https://github.com/bredelings/BAli-Phy **[Github]**

Compared to other parts of the evaluation, this stage will be fairly straightforward. In addition, for my tests, the manual for Historian, and BAli-Phy provide accurate installation instructions and can successfully be installed stock. (Historian must be built from Github).

Running Historian & BAli-Phy

To ensure that all of the tests are fair, similar parameters and settings have to be used across Historian and BAli-Phy. For instance:

- We will be using the default indel/substitution matrices for both programs
- No guiding trees will be provided for the reconstruction
- All other settings for both programs will be set to the most similar setting in the other program, or otherwise set to default.

All of the runs will be managed by a wrapper script that accounts for batching as well as the wall-clock time of each run. If run on a computer cluster, each batch can run in parallel on separate processors. However, no further speed-ups will be applied to either program to ensure authenticity of tests, and both will be run on CPU. The time can be managed through a simple time module.

The Historian run should look similar to this:

```
historian mcmc msa.fasta -v2 -upgma -preset lg
```

UPGMA is used instead of neighbor joining to enforce an ultrametric tree and support MCMC.

The BAli-phy run should look similar to this:

```
bali-phy msa.fasta -S tn93 -I rs07
```

The wrapper script should go through each of the MSAs (as found in the mapping file) and run both the Historian and BAli-phy commands one by one, and distribute batches across multiple processors. This can be done with Python's multiprocessing module.

All output files of BAli-phy and Historian should then be sorted in a similar fashion to the original sequences such that the previous mapping CSV should include all information regarding the location of the outputs.

Model Fitting for Parameters

BAli-Phy auto-generates predicted parameters (indel/substitution rate) of the evolutionary sequence within `C1.log`; however, Historian doesn't. It requires a second, `fit` command in order to generate predicted parameters. This will take longer than the reconstruction, but should also be managed by the same wrapper script to work across all sequences in batches.

```
historian fit msa.fasta -v2 -upgma > sequences.model.json
```

This will output a model file including information about indel and substitution rates which should once again be mapped back to the same CSV.

Developing Representative Tree Metrics

Key Resources (Generalized RF):

- https://doi.org/10.1093/bioinformatics/btaa614 **[Paper]**
- https://ms609.github.io/TreeDist/reference/TreeDistance.html **[Manual]**
- https://ms609.github.io/TreeDist/index.html **[Download]**

Key Resources (Align):

- https://doi.org/10.1093/bioinformatics/bti720 **[Paper]**
- https://ms609.github.io/TreeDist/reference/NyeSimilarity.html **[Manual]**
- https://ms609.github.io/TreeDist/index.html **[Download]**

Key Resources (rSPR and MCMC Mixing):

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4395846/ **[Paper]**
- https://github.com/cwhidden/rspr/ **[Github]**

The goal of the comparison between Historian's reconstructed trees to Bali-Phy's is to create a dynamic testing environment in which we highlight the strengths and weaknesses of each with Smith's generalized RF metric and Nye's "Align". A brief overview of these metrics and their characteristics are described by Kuhner and Yamato (2015). In summary, the generalized RF performs extraordinarily well in the case of similar phylogenetic trees. On the other hand, Nye's algorithm works well with non similar trees.

In addition, we will be using a modified version of rSPR (rooted subtree-prune-and-regraft) available here. This is due to SPR operators being closely related to common rearrangements in phylogenetic programs, making it ideal for analyzing MCMC behavior including mixing. More information is described by Whidden and Matsen (2015). Additionally, due to the accuracy of SPR, its implementation will provide a dynamic testing environment for evaluating Historian under various conditions.

Download and Installation

TreeDist, the required package for metric evaluation, requires a R installation on the machine. In addition, it requires `apt-get` to include CRAN on its source list. Instructions on how to do so can be found here. Afterwards, normal installation can proceed as described on TreeDist's website through the R console.

Possible Installation Issues:

1. When installing R, apt-key must be updated to add CRAN to source list and to update apt
2. RLang package must be installed for successful installation of TreeDist
3. An error may show up that cancels the install of TreeDist (Ubuntu Linux). This can be resolved with the following command:

```
sudo apt install build-essential libcurl4-gnutls-dev libxml2-dev libssl-dev
```

rSPR is not included in the TreeDist package and must be installed from the github source. This can be done by cloning the repository and typing `make`. Similarly, sprspace must be installed from this github repository.

Tree Analysis and Comparison

Using "Align" as well as the generalized RF metrics can be defined relatively simply based on the notes in the documentation. The program should simply loop through the mapping CSV and compare the ancestral tree to the reconstructed tree.

"Align"/Nye Comparison:

```
NyeSimilarity(
  tree1 = ,
  tree2 = ,
  similarity = FALSE,
  normalize = TRUE,
  normalizeMax = is.logical(normalize),
  reportMatching = FALSE,
  diag = TRUE
)
```

Generalized/Smith Comparison:

```
TreeDistance(tree1, tree2 = NULL)
```

However, after the comparison is completed, there are a number of post-processing steps that must be completed. Firstly, the similarity scores must be converted into distances. For Align, this is already done in the output. However, for the Generalized RF, a function must be written to subtract the similarity twice from the total information content such as SPI, or MSI of all the splits in both trees.

In addition, the distance metrics must be normalized to a factor of the number of nodes on the trees. The metrics can then be collected back into the same CSV as previously collected.

Unfortunately, R does not have any major libraries that supports automated parallelization of the computations and manually changing code in the TreeDistances and rSPR, which would open up a whole new can of worms. However, if necessarily required, R has a native parallelization package, and Python has an intuitive extension (Bend) that automatically parallelizes computations.

Evaluating MCMC Mixing

Evaluating MCMC Mixing in Historian and BAli-Phy will closely follow the procedures described by Whidden and Masten (2015). After installing rSPR and SPR, there are still a few config files to edit, as described in their README. After configuring sprspace, run the following commands across the pairs of reconstructed and simulated trees for each Historian and BAli-Phy.

NOTE: The following commands will likely have to be modified in order to support file formats from Historian and BAli-Phy.

```
bash process_mrbayes_posterior.bash <directory> <file.t> <file.p>
```

```
perl mean_access_time.pl [--num_trees n] [--num_trees_2 n2] [--tree_list l]
< tree_file
```

These two programs will compute mean commute time (MCT) statistics. Commute time is defined by the number of Markov chain iterations necessary to move from the highest probability topology of the given tree and back. This is used instead of mean access times (MAT) as MAT can be computationally expensive over a large data set.

In addition to this, rSPR can be utilized to calculate the rSPR distance between the simulated and reconstructed tree in addition to "Align" and Generalized RF. As stated previously, this would create a dynamic testing environment.

```
./rspr -pairwise tree1.tree tree2.tree -approx
// You can use -fpt instead of -approx for exact results
```

Data Investigation

After all of these steps have been completed, we should have a mapping file with the following information for each sequence, not necessarily in the following order. In addition, they should be in the format/units specified.

Ancestral Sequence ID	**Ancestral Sequence**	**Indel Rates**	**Substitution Rates**	**Reconstructed Historian**
Integer Value	FASTA Seq	mut/site/year	sub/site/year	NEWICK tree
Reconstructed BAli-Phy	**Wall-clock Historian**	**Wall-clock BAli-Phy**	**General RF Historian**	**"Align" Historian**
NEWICK tree	(mm:ss:ms)	(mm:ss:ms)	Norm Distance	Norm Distance
General RF BAli-Phy	**"Align" BAli-Phy**	**MCT Historian**	**MCT BAli-Phy**	**rSPR Space Historian**
Norm Distance	Norm Distance	MCT Steps	MCT Steps	Norm Distance
rSPR SPace BAli-Phy	**Retrieved Parameter Historian**	**Retrieved Parameter BAli-Phy**		
Norm Distance	mut/site/years ub/site/year	mut/site/years ub/site/year		

NOTE: Ensure that the programs (Historian, BAli-Phy, rSPR, TreeDist) utilize formats and units consistent with the above table or are later converted into the above formats.

With this information, we can do an analysis of BAli-Phy's performance to Historian. These include comparing Historian's performance (Align, Generalized RF, rSPR) to BAli-Phy across the various tests and conditions to plot trends, differences, etc. We could also perform a chi-squared test to verify the validity and significance of differences and explore MCMC mixing visually with rSPR's integration with Cytoscape. Distributions of results could be generated in respect of wall-clock time and the aforementioned metrics.

When analyzing data, some things to look out for are specific tree structures, organisms, and parameters at which Historian performs poorly compared to BAli-Phy in order for further improvement. This would go alongside looking for points of improved performance for Historian. In the end, analysis will mostly be situational depending on the final data we collect.

Aggregated References

Cory L. Strope, Stephen D. Scott, Etsuko N. Moriyama, indel-Seq-Gen: A New Protein Family Simulator Incorporating Domains, Motifs, and Indels, *Molecular Biology and Evolution*, Volume 24, Issue 3, March 2007, Pages 640–649, https://doi.org/10.1093/molbev/msl195

Sudhir Kumar, Sankar Subramanian, Mutation rates in mammalian genomes, *Proceedings of the National Academy of Sciences*, Volume 99, Issue 2, January 2002, Pages 308-808, https://doi.org/10.1073/pnas.022629899

Smeds L., Qvarnström A., Ellegren H, Direct estimate of the rate of germline mutation in a bird, *Genome Research,* Volume 26, Issue 9, September 2016, Pages 1211-1218, https://doi.org/10.1101%2Fgr.204669.116

Duffy, S., Shackelton, L. & Holmes, E. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genetics,* Volume 9, March 2008, Pages 267–276, https://doi.org/10.1038/nrg2323

Ian H Holmes, Historian: accurate reconstruction of ancestral sequences and evolutionary rates, *Bioinformatics*, Volume 33, Issue 8, April 2017, Pages 1227–1229, https://doi.org/10.1093/bioinformatics/btw791

Zardoya, R, Quest for the Best Evolutionary Model, *J Mol Evol*, Volume 89, April 2021, Pages 146-150, https://doi.org/10.1007/s00239-020-09971-z

I. Miklós, G. A. Lunter, I. Holmes, A "Long Indel" Model For Evolutionary Sequence Alignment, *Molecular Biology and Evolution*, Volume 21, Issue 3, March 2004, Pages 529–540, https://doi.org/10.1093/molbev/msh043

Benjamin D Redelings, BAli-Phy version 3: model-based co-estimation of alignment and phylogeny, *Bioinformatics*, Volume 37, Issue 18, September 2021, Pages 3032–3034, https://doi.org/10.1093/bioinformatics/btab129

Martin R Smith, Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees, *Bioinformatics*, Volume 36, Issue 20, October 2020, Pages 5007–5013, https://doi.org/10.1093/bioinformatics/btaa614

Tom M.W. Nye, Pietro Liò, Walter R. Gilks, A novel algorithm and web-based tool for comparing two alternative phylogenetic trees, *Bioinformatics*, Volume 22, Issue 1, January 2006, Pages 117–119, https://doi.org/10.1093/bioinformatics/bti720

Chris Whidden, Frederick A. Matsen, Quantifying MCMC Exploration of Phylogenetic Tree Space, *Systematic Biology*, Volume 64, Issue 3, May 2015, Pages 472–491, https://doi.org/10.1093/sysbio/syv006

Mary K. Kuhner, Jon Yamato, Practical Performance of Tree Comparison Metrics, *Systematic Biology*, Volume 64, Issue 2, March 2015, Pages 205–214, https://doi.org/10.1093/sysbio/syu085