Phyloage User Guide

Dirk Struve
phylofriend at projectory.de
https://github.com/yogischogi/phyloage/

July 31, 2016

Contents

1 Introduction

Phyloage is an experimental program for genetic genealogy and history research. It combines SNP and Y-STR results to calculate TMRCA (Time To Most Recent Common Ancestor) estimates from STR mutational differences.

Since the emergence of next generation sequencing, SNP based phylogenetic trees are rapidly evolving and less attention is paid to Y-STR mutations. But if we look at the properties of SNP mutations in detail, we see that they are not without drawbacks:

- SNP mutations are extremely stable. Phylogenetic trees based on SNPs are highly reliable.

- Not all SNPs are useful for genealogical purposes [1]. This is problematic for TMRCA calculations because wrong SNPs lead to false results.

- For TMRCA calculations the margins of error are rather large. YFull has tried to include only genealgical relevant SNPs into their phylogenetic tree [11] and estimates one mutation every 140 years [1].

It would be nice to have more precise estimates and a second independent method for TMRCA calculations, that can be used to verify and complement the results gained by SNP counting. This could be achieved by using the well known Y-STR mutations. Their properties are rather different from SNPs:

- STRs have been used for genetic genealogy for a long time. Some specific marker sets are well known and many people have tested.

- Next generation sequencing reveals results for up to 500 markers. This should give us enhanced precision for TMRCA estimates.

- Time estimates using STRs are often biased towards specific lineages that proliferated more than others.

- STRs mutate back and forth. It is impossible to create reliable phylogenetic trees that go back deep into history. Furthermore the back and forth mutations lead to a saturation effect that invalidates time estimates for long time spans (thousands of years).

There have been numerous efforts to overcome the disadvantages of STR counting. A small overview is given in [3].

Phyloage introduces a new approach. The basic idea is to start with an SNP based phylogenetic tree, insert the Y-STR results and calculate modal haplotypes for each node of the tree.

In theory this should reduce saturation effects and give better time estimates for both deep history and genealogical time frames.

If you use this program, please remember that it is still highly experimental. At the time I am writing this (March 1, 2016), there is no person on earth who has a long time experience with 500 Y-STR markers.

Have fun experimenting!

Dirk

2 Installation

This guide is mainly targeted towards persons who use Linux Mint or other Linux versions of the Debian family. Some familiarity with the use of Linux commands is assumed.

Currently there are no binary distributions available for Windows or the Mac. Users of these operating systems can use Phyloage as well, but they will experience some laborious installation work. The best way is to follow the instructions provided on the Go home page.

The following list applies to Linux users only:

1. Make sure that the Go programming language is installed. If not it can be installed by typing
   `sudo apt-get install golang`

2. Read the Go Getting Started guide. Make sure to set your *GOPATH* variable and include it in your *PATH* so that Go programs can be found.

3. Fetch the Phyloage program with
   `go get github.com/yogischogi/phyloage`

4. Install the program with
   `go install github.com/yogischogi/phyloage`

You should end up with two newly installed program, Phyloage and Phylofriend [8]. Phylofriend does a lot of the background calculations for Phyloage.

3   Command Line Options

Command line options may be given in arbitrary order. Parameters may be specified by using a space or equals sign. For example the following options are identical: `-treein=mytree`, `-treein mytree`.

-help Prints available program options.

-treein Filename of the SNP based phylogenetic tree.

-treeout Filename of the results tree in text format.

-personsin Filename or directory of files containing the persons' Y-STR values. If this is a single file it must contain results for multiple persons. The input file format is CSV (comma separated values) or text format.

If a directory is provided for input it must contain multiple files in YFull format, each file containing the results for a single person. The person's ID is extracted from the filename.

`personsin` supports multiple file names separated by commas.

-mrin Filename of the mutation rates to use.

-gentime Generation time.

-cal Calibration factor.

-offset An offset that is added to all calculated ages.

-subclade Selects a branch of the tree specified by an SNP.

-htmlout Output filename for Y-STR markers in HTML format.

-statistics Prints out marker statistics.

-inspect Prints out details about the specified SNPs or sample IDs. The search terms must be specified by a comma separated list, for example `-inspect=CTS4528,S11481,S14328`.

-trace Prints out a phylogenetic tree that contains the mutational values for the specified Y-STR markers. Example: `-trace=DYS393,DYS19`.

-method Method to be used for calculating modal haplotypes: `phylofriend` or `parsimony`. The default method is `parsimony`, which uses a maximum parsimony algorithm.

-stage Processing stage for the parsimony algorithm. This should be used for debugging or to see in detail what the algorithm does. The following stages are valid:

1 Calculates haplotype values that strictly satisfy the maximum parsimony criterion.

2 Calculates average haplotype values for the rest by using real numbers.

3 Replaces real numbers by real world mutation values and sets all uncertain values to -1. This stage is only for visualization and debugging.

4 Forces all values to real world mutation values.

4 Examples

4.1 Using YFull Results

SNP Input Tree

Before you can start you need to create a phylogenetic tree, that contains SNPs and the IDs of the genetic samples. The file format is text based and looks like this:

```
// This is an example tree.
// Comments begin with //.

CTS4528, S1200
    S11481
        id:YF01234
        id:YF00301
        id:YF02016
    S14328
        id:YF04242
        id:YF00101
        id:YF01010
```

Each line of the tree contains one or more SNPs or a sample ID. Subclades and samples are indented by using tabs. Each sample starts with `id:` followed by the ID. In our case these are typical YFull IDs but Phyloage supports Family Tree DNA data as well. Phyloage uses the Phylofriend [8] program for data import and many calculations as well. See the Phylofriend User Guide [9] for the details of the supported input formats.

To create a results tree:

1. Save the input tree to a file, for example *tree.txt*.

2. Download the Y-STR results for the samples from YFull and put them into a separate directory, for example *allsamples*

3. Execute the following command from a command line (the file *111-average.txt* can be found in the *mutationrates* directory of the Phylofriend program [8]):
   ```
   phyloage -treein tree.txt -treeout results.txt
   -personsin allsamples -mrin 111-average.txt -gentime 32
   ```

The results will be stored in a file named *results.txt*. We have used average mutation rates for 111 markers and a generation time of 32 years.

Results Tree

Now the file *results.txt* contains a tree with TMRCA estimates that looks like this:

```
CTS4528, S1200, STRs Downstream: 145, formed: 4644, TMRCA: 4644
    S11481, STR-Count: 140, STRs Downstream: 51, formed: 6141, TMRCA: 1647
        id:YF01234, STR-Count: 30
        id:YF00301, STR-Count: 73
        id:YF02016, STR-Count: 52
    S14328, STR-Count: 17, STRs Downstream: 81, formed: 3148, TMRCA: 2597
        id:YF04242, STR-Count: 72
        id:YF00101, STR-Count: 91
        id:YF01010, STR-Count: 82
```

Because we have used mutation rates for 111 markers, that were calibrated by using generations, the `STR count` is given in generations. It says for how long a Clade has existed before it developed any subclades. The `STRs Downstream` is a measure for the TMRCA.

`formed` denotes the age of the clade in years. It is calculated by adding `STR count` to `STRs Downstream` and multiplying the result by the generation time (the `gentime` parameter) and an additional calibration factor (1 by default).

`TMRCA` is the same as `STRs Downstream`. It is just multiplied by the generation time and the calibration factor.

4.2   Mutation Counting on the 500 Marker Scale

In the previous example we have used the average mutation rates for 111 markers. It is a good idea to start with those markers because they are well known and have been used for years.

YFull on the other hand reports up to 500 STR markers. Most persons get about 400 values out of their Big Y test. So we like to use them, but keep in mind that we do not have a long time experience with the 500 marker scale.

The input tree will be exactly the same as before but this time we execute the command:

```
phyloage -treein tree.txt -treeout results.txt
-personsin=allsample -mrin=500-count.txt -cal=39
```

The above command will do mutation counting using all markers available and then upscale the results to 500 markers. The `500-count.txt` contains

the mutation rates for marker counting. It can be found in the Phylofriend [8] *mutationrates* directory.

Currently it seems like one mutation counts as 39 years using Phylofriend's mutation model. So a calibration factor of 39 is used.

4.3 Adding Family Tree DNA Results

It is possible to add Family Tree DNA results to the input tree. This is very useful if some persons only did test on the 67 or 111 marker scale. It is also possible to add 12 or 37 marker results but the average mutation rates are different and the margins of error are very high. So I do not recommend it.

To add Family Tree DNA results, just insert their IDs (kit numbers) into the input tree. In this example we add 12345 and 67890 to the input tree:

```
CTS4528, S1200
    S11481
        id:YF01234
        id:YF00301
        id:YF02016
        id:12345
        id:67890
    S14328
        id:YF04242
        id:YF00101
        id:YF01010
```

To create the results tree:

1. Save the input tree to the file *tree.txt*.

2. Put the YFull results into a directory named *yfull*.

3. Save the Family Tree DNA results in a spreadsheet called *cts4528.csv*. The first column of the spreadsheet must contain the Family Tree DNA kit numbers. For additional details about the file format, please consult the Phylofriend User Guide [9].

4. Execute the following command from a command line:
   ```
   phyloage -treein tree.txt -treeout results.txt
   -personsin yfull,cts4528.csv -mrin 111-average.txt
   -gentime 32
   ```

The results can be found in the file *results.txt*. Of course it is also possible to build a tree only from Family Tree DNA samples and completely leave out YFull results.

4.4   Pure Mutation Counting

If you do not have files containing detailed genetic results, but you know the number of the individual (private) STR mutations, it is possible to use the input tree for counting. Just add the STR numbers to the tree like this (you can also use this method for SNP counting):

```
CTS4528, S1200
    S11481, STR-Count: 2
        id:YF01234, STR-Count: 30
        id:YF00301, STR-Count: 40
        id:YF02016, STR-Count: 50
    S14328, STR-Count: 2
        id:YF04242, STR-Count: 30
        id:YF00101, STR-Count: 40
        id:YF01010, STR-Count: 50
```

In this example we assume that one mutation counts as 100 years. Thus we use a calibration factor of 100. To create the result tree, type:

```
phyloage -treein tree.txt -treeout results.txt -cal 100
```

The result tree contains TMRCA estimates for all clades:

```
CTS4528, S1200, STRs Downstream: 42, formed: 4200, TMRCA: 4200
    S11481, STR-Count: 2, STRs Downstream: 40, formed: 4200, TMRCA: 4000
        id:YF01234, STR-Count: 30
        id:YF00301, STR-Count: 40
        id:YF02016, STR-Count: 50
    S14328, STR-Count: 2, STRs Downstream: 40, formed: 4200, TMRCA: 4000
        id:YF04242, STR-Count: 30
        id:YF00101, STR-Count: 40
        id:YF01010, STR-Count: 50
```

5   Theory

5.1   Modal Haplotype

Phyloage uses Y-STR mutation counting to estimate the TMRCA (Time to
Most Recent Common Ancestor) for a group of persons. This is not an easy
task because STR markers jump back and forth. If a marker has changed
from one value to another and back again we do not know if a mutation has
taken place or not. So for long time spans STR mutation counting does not
work and we get wrong TMRCA estimates. In many cases this effect can be
calculated and compensated as it has been done in [5]. However this method
has it's limits.

   Phyloage uses another approach. It combines SNP based phylogenetic
trees and Y-STR counting. Because of the use of SNPs, the ancestral po-
sitions on the phylogenetic tree are well known. Phyloage tries to calculate
the ancestral haplotype for each ancestor and uses Y-STR mutation counting
between those haplotypes and the modern samples. The calculated haplo-
types are not necessarily identical to the real ancestral values. So they are
called modal haplotypes.

   The calculation of a modal haplotype is not always possible. Figure 1
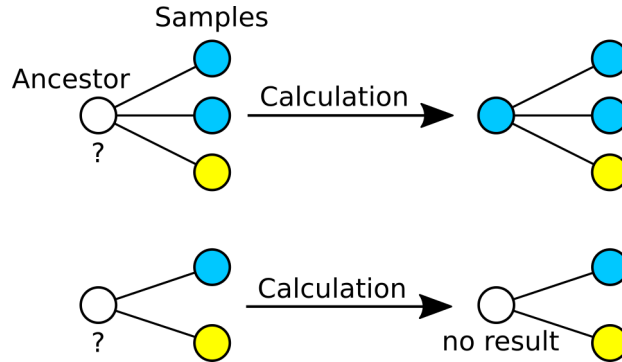illustrates the situation.

Figure 1:   Calculation of an ancestor's mutational value from a set of sam-
ples. The calculation is only possible if we have at least three samples.

   Because mutation rates are usually very slow we may assume that a mu-
tation occurs rarely. In this case the modal value of a mutation is identical to
the value that occurs most often among the samples. The rare cases are the
mutations. If the time spans get very long this is no longer true. We have
no idea how many mutations have occurred between the ancestral haplotype
and the modern samples. The mutational values are just random. If we use

them to calculate the modal haplotype, the result will be different from the real ancestral values.

So for a valid calculation of a modal haplotype, two conditions must be satisfied:

1. We must have at least three samples and two of them must have the same mutational value.

2. The time span between two ancestral haplotypes must be small compared to the average time in which a mutation occurs.

5.2   Maximum Parsimony Method

As default Phyloage uses an algorithm that tries to satisfy the maximum parsimony criterion [10]. This means that the tree that contains the least amount of mutations is considered best.

Although this is very intuitive and widely used, it is not always true or easy to calculate because

1. In many cases there is no best tree, but a number of viable solutions. As an easy example consider two persons with two different values for the same marker. Both values are valid solutions for the modal haplotype.

2. Long time single lineages may introduce a set of completely random values, because it is likely that marker's have changed several times. The criterion of maximum parsimony can be misleading [2].

3. The calculation of the most parsimonious tree can require enormous amounts of computing time [10]. In such cases we prefer a good solution over the best one possible.
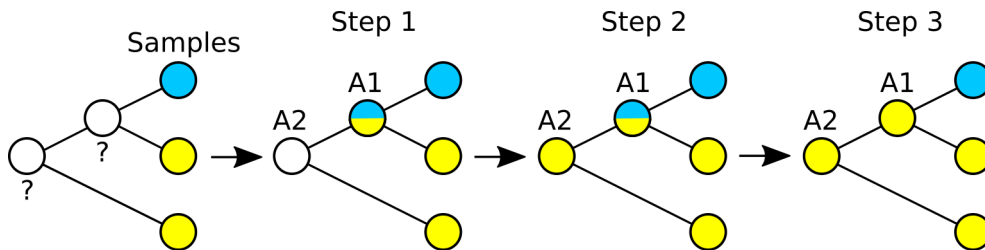

Figure 2:   An algorithm that satisfies the criterion of maximum parsimony. The tree is calculated backwards until a unique modal value is found. After that the missing values are recalculated.

Figure 2 shows how the most parsimonious solution is derived for a very simple phylogenetic tree. In this example the modal values are calculated backwards in time until a valid solution is found. After that the tree is recalculated from the past to the present to find modal values for nodes that did not have a viable solution before.

In reality the situation is more complicated because in many cases there simply is no valid solution for a certain marker. The algorithm that is used by Phyloage works like this:

1. If possible, modal values that strictly satisfy the maximum parsimony criterion are calculated. These are the easy cases.

2. The rest of the modal values are calculated by using averages and real numbers. This solution should come close to the maximum parsimony criterion, at least for many samples. In reality this solution is not possible because real marker values are restricted to whole numbers.

3. The modal values are mapped to their nearest neighbors from the set of real marker values. Still some markers can not be calculated because they do not have a unique nearest neighbor.

4. All markers that could not be calculated in the top node are forced to a real world neighbor value. After that the whole tree is recalculated and all markers are forced to a real world value.

The algorithm does not necessarily find the best solution possible, but it should find a good one. For more information please consult the program's source code documentation [7] directly.

5.3 Phylofriend Method

Phyloage also provides an alternative method to calculate the modal haplotypes. It is very simple and uses the same algorithm as the Phylofriend [8] program.

Figure 3 illustrates how Phylofriend calculates a modal haplotype. Generally it chooses the most common value as the ancestral value. If there is no most common value it just guesses among the valid ones. This works very well if a modal haplotype has many samples (at least three) downstream of it's own. For just two samples the resulting modal haplotype will be somewhere in between the two sample haplotypes.

This algorithm does not satisfy the maximum parsimony criterion, but for sparse trees it works better than some might expect. The reason is that a
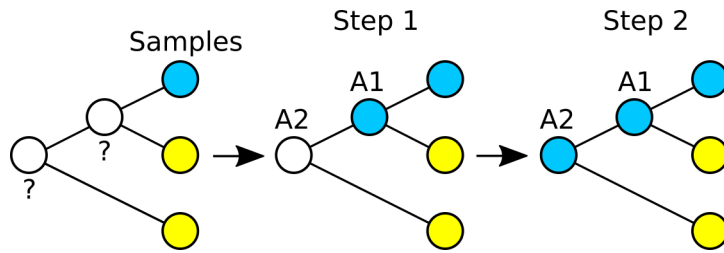
Figure 3: The Phylofriend method to determine a modal haplotype is very simple. If a modal value can't be calculated, it just guesses and continues the calculation.

modal haplotype, that is calculated only from samples, can not be influenced by older modal values afterwards, thus eliminating the effect of distant single lineages.

What method to use depends highly on tree structure and research goals. I am afraid that I can not give a general recommendation. For sparsely populated trees both methods won't work very well due to the difficulties involved. For densely populated trees both methods, maximum parsimony and phylofriend, should yield similar results.

References

[1] Dmitry Adamov, Vladimir Guryanov, Sergey Karzhavin, Vladimir Tagankin, Vadim Urasin. *Defining a New Rate Constant for Y-Chromosome SNPs based on Full Sequencing Data*. The Russian Journal of Genetic Genealogy (Русская версия), Vol 6, No 2 (2014)/Vol 7, No 1 (2015).

[2] Joseph Felsenstein, *Cases in which Parsimony or Compatibility Methods will be Positively Misleading*, Systematic Zoology 27 (4): 401–410, 1978, doi: 10.1093/sysbio/27.4.401.

[3] David Hamilton, *An accurate genetic clock*, bioRxiv preprint, first posted online June 15, 2015, doi: 10.1101/020933.

[4] ISOGG, *Y-DNA Haplogroup R and its Subclades*. Date visited: 2015-10-05.

[5] Anatole A. Klyosov, *DNA Genealogy, Mutation Rates, and Some Historical Evidence Written in Y-Chromosome, Part I: Basic Principles and the Method*. Journal of Genetic Genealogy, 5(2):186-216, 2009.

[6] Sergey Malyshev, *R1b-M269 (P312- U106-) DNA Project (aka ht35 Project) Phylogenetic Tree*. R1b-M269 (P312- U106-) DNA Project, Date visited: 2015-10-05.

[7] Dirk Struve, *Phyloage source code documentation*, GoDoc, 2016.

[8] Dirk Struve, *Phylofriend, a program to calculate genetic distances*. Google Project Hosting, 2014; GitHub, 2015.

[9] Dirk Struve, *Phylofriend User Guide*. Google Project Hosting, 2014; GitHub, 2015.

[10] Wikipedia, *Maximum parsimony*. Date accessed: 2016-03-28.

[11] YFull, *YFull Phylogenetic Tree*. Date visited: 2016-03-01.