Phylofriend User Guide

Dirk Struve
phylofriend at projectory.de
https://code.google.com/p/phylofriend/

September 4, 2014

Contents

1 Introduction

Phylofriend's main purpose is to calculate genetic distances from Y-DNA data. The results can be used as input for the PHYLIP[3] program to create phylogenetic trees.

When I started creating phylogenetic trees I often found myself in a difficult position. As a Linux user I was missing some of the tools available under Windows. So I started to write this program to fill in the gaps and make myself comfortable again.

This does not mean that you can not use Phylofriend when working under Windows or the Mac. But currently there is no binary distribution available and you will probably face a hard time installing Phylofriend and the associated programs. So I only recommend this if you are an experienced user.

Phylofriend has some nice features. It can be used

- to create phylogenetic trees using the PHYLIP[3] program. Y-STR values from Family Tree DNA projects can easily be imported.
- as a programming library. Phylofriend is written in Google's Go programming language. This language is not only suited to solve Google's large scale programming problems. It is also an excellent tool for part time programmers who have to concentrate on their projects (often students).
- to extract Y-DNA data from Family Tree DNA projects and convert it into simpler text files that are better suited for further processing.
- to automate phylogenetic tree creation. Phylofriend is a command line tool and this scares many people away. But if you have to repeat the same tasks over and over again you will eventually start to write some scripts and this is where command line tools come in handy.

I hope this program will be useful. Have a good time!

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 Phylofriend as well, but they will experience some laborious installation work. The best way is to follow the instructions provided on the Go home page and the PHYLIP 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. For the creation of phylogenetic trees install the PHYLIP program package by typing
   `sudo apt-get install phylip`

4. Fetch the Phylofriend program with
   `go get code.google.com/p/phylofriend`

5. Install the program with
   `go install code.google.com/p/phylofriend`

3 Command Line Options

Command line options may be given in arbitrary order.

-help Prints available program options.

-personsin Filename of the file containing the persons' Y-STR values. This may be in CSV (comma separated values) or text format.

-namescol Number of the column that contains the names when reading CSV files.

-mrin Filename of the mutation rates to use.

-anonymize If this is true persons' names are replaced by numbers.

-modal Creates modal haplotype.

-phylipout Filename for the distance matrix that can be fed into the PHYLIP[3] program.

-mrout Filename for the output of the currently used mutation rates.

-txtout Filename for text output of persons and Y-STR values.

-nvalues Number of Y-STR values to write to the text output file.

-gendist Generation distance.

-cal Calibration factor.

4 Examples

4.1 Create a Phylogenetic Tree

1. Copy persons' data from a Family Tree DNA project website into a spreadsheet. If the Y-STR values do not appear properly try inserting them into the spreadsheet as unformatted text.

2. Save the spreadsheet in CSV (comma separated values) format, for example *persons.csv*.

3. Start a terminal or command line interpreter and go to the directory where you stored *persons.csv*.

4. Create a matrix of genetic distances by typing
   `phylofriend -personsin persons.csv -phylipout infile`

5. Use the PHYLIP program to create a tree in Newick format with
   `/usr/lib/phylip/bin/kitsch`
   You will need to answer some questions. Usually the default values are good enough. The results will be two text files, one named *outtree* which contains the tree in Newick format and another one named *outfile* which contains a more human readable description.

6. Create an image of the tree by typing
   `/usr/lib/phylip/bin/drawgram`
   Use *outtree* as the input file name. The resulting tree will be stored in a file named *plotfile*.

   A nice alternative to visualize the tree is the use of the Trex[6] web-server. You can copy the contents of the file *outtree* into the Trex window.

If you do not specify a file containing mutation rates Phylofriend will use average 37 marker values as default. They can be found in [5].

4.2 Pimp Your Tree with Nicer Labels

By default Phylofriend assumes that your persons input file's first column contains a list of IDs. This is usually a Family Tree DNA Kit number. The resulting tree is hard to read. Many projects keep names in another column. You can access this column with the *namescol* option. Suppose your second column contains names. You can create a distance matrix with names instead of IDs by typing

```
phylofriend -personsin persons.csv -namescol 2
-phylipout infile
```

Due to compatibility issues with other programs the names should be 10 characters long and may not contain Unicode characters. Phylofriend will apply a name transformation but the result is sometimes a bit strange.

You can also use the *namescol* option to create trees that contain the origins of people or the haplogroups. Although I strongly recommend to build trees only from people who belong to the same haplogroup this is sometimes useful if you want to know if different haplogroups are close on their Y-STR values.

If you want to publish your tree you will often need to protect the privacy of the members. This is what the *anonymize* option is for. By typing

```
phylofriend -personsin persons.csv -phylipout infile -anonymize
```

you will get a distance matrix where the names are replaced by numbers.

4.3 Use a Specific Set of Mutation Rates

The first example uses Phylofriend's build in mutation rates which are average values for the standard 37 marker test. Phylofriend supports the use of arbitrary mutation rates by the *mrfile* option. The *phylofriend/mutationrates* directory contains some files with mutation rates. The average mutation rates where taken from [5]. If you like to compare on 67 markers you can use

```
phylofriend -personsin persons.csv -phylipout infile
-mrin code.google.com/p/phylofriend/mutationrates/67-average.txt
```

4.4 Calibrate Your Data

Mutation rates depend on the method applied to calculate genetic distances and the sample populations used. Mutations themselves occur by coincidence. Average mutation rates often yield acceptable results but in most cases you will have to calibrate your data especially if you want to calculate in years.

Phylofriend provides two options for data calibration: *gendist*, the generation distance in years and *cal* an additional calibration factor. Internally they are just multiplied together but using two separate factors seems more convenient for typical use cases.

The default value for the generation distance is 25 years. For time spans over the last few hundred years a generation distance of 30 years often yields better results. This can be done by

```
phylofriend -personsin persons.csv -phylipout infile
-gendist 30
```

It is often difficult to calibrate data because you need a reliable paper trail or a well defined historic event. If you are lacking both you can try to apply Klyosov's statistical method[4]. For large enough sample sizes this will effectively reduce the statistical error but you will still be left with an unknown systematical one.

4.5 Count Mutations

The *phylofriend/mutationrates* directory contains sets of mutation rates where all markers are set to 1. This makes mutation counting easy but you will need an additional little trick. Internally Phylofriend uses average values. So it is possible to compare persons who have tested on different sets of markers.

If you want to count mutations for example on a 37 marker scale, you must multiply Phylofriend's internal results with 37. The easiest way to archive this is by misusing the *gendist* option.

```
phylofriend -personsin persons.csv
-mrin code.google.com/p/phylofriend/mutationrates/37-1.txt
-phylipout distancecount.txt -gendist 37
```

4.6 Extract Data from a Spreadsheet

Spreadsheet data is often uncomfortable to handle, especially if you want to write your own program and need to parse it. For this purpose Phylofriend supplies the *txtout* option. It writes data to a text file in simplified form.

The easiest way to use it is

```
phylofriend -personsin persons.csv -txtout persons.txt
```

This extracts the data from *persons.csv* and writes it to *persons.txt*. The first column of *persons.txt* contains the first column found in *persons.csv*, usually a set of IDs. The following columns contain the Y-STR values. All columns are separated by tabs.

If you want to use another column as the first column you can use the *namescol* option:

```
phylofriend -personsin persons.csv -txtout persons.txt
-namescol 2
```

This extracts the second column from *persons.csv* and writes it to the first column of *persons.txt*.

When using the *txtout* option you will need to specify how many Y-STR values are written to the text file. This is done by *nvalues*. Phylofriend will write the same number of Y-STR values to each line. Missing values are written as 0, larger sets of Y-STR values are truncated. This is how to write a full set of 111 markers:

```
phylofriend -personsin persons.csv -txtout persons.txt
-nvalues 111
```

5 Technicalities

5.1 Source Code Documentation

To access the source code documentation point your web browser to:

- http://godoc.org/code.google.com/p/phylofriend

If you want to modify the source code it is best to use *godoc* locally on your computer.

1. If *godoc* is not yet installed install it by typing
   `sudo apt-get install golang-go.tools`
   (Linux Mint)
2. Start the documentation web server with
   `godoc -http=:6060`
3. Point your web browser to `localhost:6060`.
4. Click on *Packages* and search for *phylofriend*.

This will give you a nice overview of the internal program documentation. You can also click on function names to browse the source code.

5.2 Mutation Model

The two basic mutation models are the infinite alleles model and the stepwise mutation model as explained by Bruce Walsh[8]. Phylofriend uses a hybrid mutation model. Most markers are calculated using the stepwise model, palindromic markers are calculated as described in [2].

As the method for calculating the genetic distance is likely to change with time please look at the internal program documentation if you need more details.

5.3 Mutation Rates

The directory *phylofriend/mutationrates* contains files with predefined mutation rates.

All 1 Mutation Rates

- 12-1.txt
- 37-1.txt
- 67-1.txt
- 111-1.txt

In these files all mutation rates are set to 1 for the corresponding set of markers. They are useful for simple mutation counting.

Average Mutation Rates

- 12-average.txt
- 37-average.txt
- 67-average.txt
- 111-average.txt

In these files average mutation rates are used for the corresponding set of markers. The mutation rates were taken from [5]. The distance matrix contains years but these years should not be interpreted as TMRCA (Time to Most Recent Common Ancestor). It is more like a *How long do I have to wait on average before such a genetic distance occurs?*

These mutation rates are useful for genealogical purposes. During genealogical time frames (about 500 years) long time stable markers rarely change. If such an event happens a son may be immediately 1000 years away from his father in terms of genetic distance. This is a correct result because mutations are governed by coincidence and such events do happen from time to time but it does not make much sense to place a son far away from his family. So very stable markers are purposely underweighted by using average values.

12 Marker Single Mutation Rates

- 12.txt
- 37-12.txt
- 67-12.txt

- 111-12.txt

These files are basically the same as the average mutation rates files but for the first 12 markers the mutation rates are set per marker. This puts on the appropriate weight on very stable markers. The mutation rates were taken from Wikipedia[7]. Although there may be some doubt if data from Wikipedia can be trusted the first 12 markers are well known and they are in use for a long time. So I just adopted them. The values should be good enough for most purposes.

For each file the 12 marker mutation rates were multiplied by a calibration factor so that their average value matches that from the rest of the file.

These files are useful for deeper history or if you observe changes on long time stable markers. If several persons share the same value on a very stable marker they probably belong together. So try these mutation rates to see if the results make more sense than those obtained by using average markers.

5.4 CSV Input Format

Example:

```
id1,"Dirk Struve",Germany,R1b-CTS4528,13,24,14,11,11-14,12,...
id2,"Pyl. O. Friend",Germany,R1b-CTS4528,13,24,14,11,11-14,...
```

When importing a file in comma separated values format the first column must contain IDs. An arbitrary number of columns containing custom information may follow. The last columns must contain at least 12 Y-STR values in Family Tree DNA order. Rows containing comments are allowed.

Phylofriend will always try to parse the file as best as it can.

5.5 Text Format

Example:

```
Dirk_Struv 13 24 14 11 11 14 12 12 12 12 14 28
Pyl._O._Fr 13 24 14 11 11 14 12 12 12 12 14 28
```

The text format is a simplified format intended for easy parsing and to work well with other programs. For compatibility reasons the first column is exactly 10 characters long and contains only non Unicode characters. Spaces are transformed into underscores. The following columns contain Y-STR values separated by tabs.

5.6 PHYLIP Format

Example:

```
2
Dirk_Struv 0 0
Pyl._O._Fr 0 0
```

The first line contains the number of entries. An entry line contains an ID that is 10 characters long and contains only non Unicode characters. Spaces are transformed into underscores. The columns containing genetic distances are separated by tabs. For readability reasons Phylofriend writes only integers. If you need more precision you can scale the distance by using the *cal* option.

6 Theory

6.1 Haplogroups

The human Y chromosome is only inherited from father to son. Usually large parts of it are unchanged but sometimes a mutation at a single position occurs. Such mutations are called SNPs (Single Nucleotide Polymorphisms). They are very stable and can be used to group people into ancestral lines.

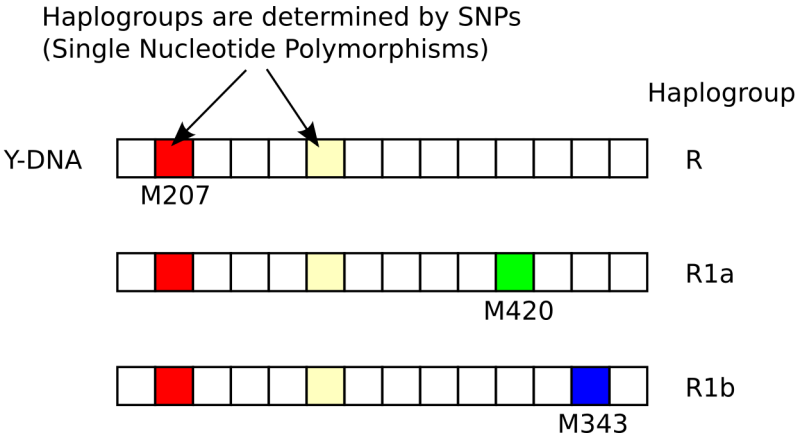

Figure 1: A set of single mutations defines a person's haplogroup. Persons who belong to the same haplogroup usually share a common ancestor within a few thousand years.

Figure 1 illustrates a typical situation. Many Europeans belong to haplogroup R, which is characterized by the mutation M207 and others which are not mentioned here. The M207 marker is inherited throughout all following ancestral lines. Later the mutations M420 and M343 occurred. They are mutually exclusive to each other thus defining separate lineages. The marker M420 defines the haplogroup R1a which is common in Eastern Europe and the marker M343 defines the haplogroup R1b which is common in Western Europe. Because all people who belong to R1a and R1b share the M207 marker we know that they also share a common ancestor a long time ago.

SNPs do not occur very often. Persons who belong to the same haplogroup usually share a common ancestor within a few thousand years. If you know your haplogroup it will tell you something about your deep history (thousands of years).

6.2 Haplotypes

Most people want to know more about family relationships. Luckily there is another type of mutations that occurs much more often than SNPs. This is the repetition of certain genetic patterns. They are called STRs (Short Tandem Repeats) and group people into haplotypes.

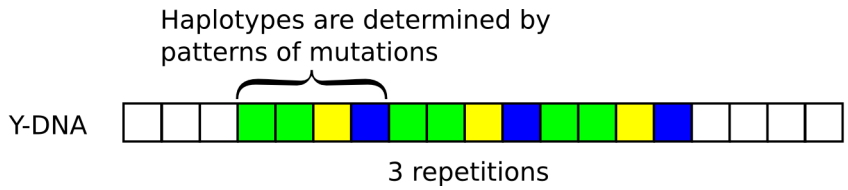

Figure 2: A person's haplotype is defined by patterns of mutations. Persons with the same haplotype usually share a common ancestor within a few hundred years if they take a standard test with 37 markers or more.

On the Y chromosome there are many different patterns that repeat themselves. Every pattern has a name, for example DYS393. If we count the number of repetitions for a specific pattern we get a value, for example DYS393=13. Figure 2 illustrates the situation for an exemplary marker. The pattern is repeated 3 times. So this marker has a value of 3.

If a mutation occurs the number of repetitions changes. A single marker mutates rarely but the modern standard tests use many markers at the same time, most often 37, 67 or 111. The combination of all these markers defines a haplotype. Persons who share the same haplotype usually also share a common ancestor within a few hundred years.

It is important to realize that haplotypes are not as good as haplogroups. Haplotypes sometimes overlap between separate lineages. So when working with haplotypes the haplogroup should also be known.

Phylofriend uses haplotypes to calculate genetic distances between persons by counting the number of different mutations.

6.3 Phylogenetic Trees

Phylogenetic trees group people together according to their genetic distances. They give us a picture of how different persons are related.

Figure 3 shows some exemplary trees. If a father has two sons we would expect that both sons are at a little genetic distance away from their father. If they are not twins they also should be separated by each other. Tree a illustrates this situation. This is exactly the tree we would get if we could measure the genetic distance at a high resolution and know the birth dates

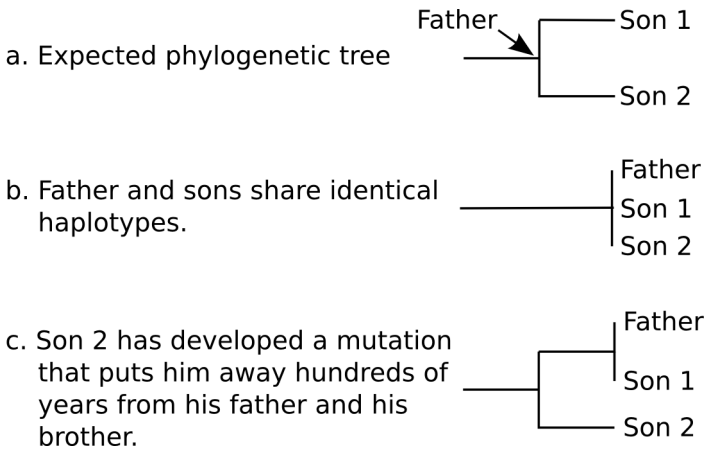

Figure 3: Phylogenetic trees group persons together according to their genetic distances. Genetic distances are often associated with time scales but this is only true for long time spans.

of father and sons. Because we know the birth dates we would put the father before the sons and associate the horizontal axis with a time scale assuming that genetic distance is proportional to time.

Reality however shows a different picture. The currently available standard tests (37, 67 and 111 markers) only offer a low resolution. So most often we can not measure any genetic distance between father and sons. This is illustrated by tree b. Because the tree illustrates genetic distances and not time distances father and sons are all side by side.

If son 2 develops a mutation by chance we get a confusing picture. Son 2 is suddenly far away from his father and his brother. This is shown by tree c. The reason for the great distance between father and son 2 is the low resolution of the test. If we take the mutation rates from [5] we can calculate how long it takes on average for a mutation to occur. For a 37 marker test the result is 280 years, for a 67 marker test 210 years and for a 111 marker test 130 years.

This means if the father and his sons took a 37 marker test and son 2 has developed a mutation by chance he will appear to be 280 years away from his father but the only reason for that is the low resolution of the test.

In most cases the standard tests are good enough. We do not need a paternity test for genealogical purposes but it is important to remember that the standard tests only places persons into time frames of hundreds of years.

6.4 Genetic Distances

There are many ways to calculate genetic distances. Here we use the method of mutation counting as described by Anatole Klyosov in [4]. We just count the number of mutation one persons differs from another and take the result as the genetic distance. Take a look at these two 6-marker haplotypes:

	DYS393	DYS390	DYS19	DYS391	DYS385a	DYS385b
Carl	13	24	14	11	11	14
Clas	13	**23**	14	**12**	11	14

At DYS390 and DYS391 Clas differs from Carl by one mutation. Thus their genetic distance on the 6-marker scale is 2.

To represent this genetic distance in years we need to know how often mutations occur. Thus we need a mutation rate. The mutation rate is defined as follows:

$$k = \frac{m}{gY} \tag{1}$$

k: Mutation rate per marker and generation
m: Number of mutations
g: Number of generations
Y: Number of Y-STR values (markers)

Mutation rates are derived from sample populations. Different family lineages often expose different mutation rates. So the standard mutation rates should be considered as a first estimate. You can uses Phylofriend's *cal* option to adjust your data to historic events.

Now we can use formula 1 to calculate a genetic distance in generations. Formula 1 is equivalent to:

$$g = \frac{m}{kY} \tag{2}$$

This gives us the number of generations between Carl and Clas. The only thing we still need to know is the generation time. For historical purposes a generation time of 25 years is commonly used. We get the time by multiplying the number of generations with the generation time.

$$t = gd \tag{3}$$

t: Genetic distance in years
g: Number of generations
d: Generation time in years

By substituting g with formula 2 we get the genetic distance in years:

$$t = \frac{md}{kY} \tag{4}$$

t: Genetic distance in years
m: Number of mutations
d: Generation time in years
k: Mutation rate per marker and generation
Y: Number of Y-STR values (markers)

Let us try this out. The number of mutations between Clas and Carl is 2. The mutation rate for a 6-marker haplotype is 0.002 [4]. So

$$t = \frac{2 \cdot 25}{0.002 \cdot 6} = 4167 years \tag{5}$$

Wow, that is a very long time. But what does this number actually mean? First, it is an average value and there is a very big margin of error to it but it is useful as a first estimate and to group people together according to their genetic distance. Second it is the time Clas and Carl would be separated if Clas would be an ancestor of Carl (or the other way round). It is not the time to their most recent common ancestor (TMRCA).

If Clas and Carl would be living today, how long would be the the time to their most recent common ancestor? The first guess is that the common ancestor would be in-between his descendants in terms of genetic distance. So Claus and Carl would be $4167 years/2 = 2085 years$ away from him.

This is a good first guess but unfortunately reality is much more complicated. Mutations occur by coincidence and due to the laws of statistics some people develop only a few mutations while others develop much of them. So our first guess may give a totally wrong impression. Generally it is not a good idea to calculate the time to a common ancestor just based on the results of two people. The best way is to identify a group of people who share a common lineage and calculate the time to the most recent common ancestor for the whole group. Anatole Klyosov describes how this is done in [4].

For those who still want to use a TMRCA value based on the results of just two people Bruce Walsh has developed a method to give a time estimate[9]. This is better than the naive calculation we have used before but it is still an estimate and it is only valid for demographically stable populations.

References

[1] James Archie, William H.E. Day, Wayne Maddison, Christopher Meacham, F. James Rohlf, David Swofford, Joe Felsenstein, *The Newick Tree Format*. 1986, Date accessed: 2014-08-06.

[2] R. A. Canada, *How does the infinite allele comparison method work for palindromic markers?*. Family Tree DNA, 2014, Date accessed: 2014-08-07.

[3] Joe Felsenstein, *PHYLIP, a free package of programs for inferring phylogenies.*. University of Washington, First version: 1980, Date accessed: 2014-08-06.

[4] 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.

[5] Anatole A. Klyosov, *Ancient History of the Arbins, Bearers of Haplogroup R1b, from Central Asia to Europe, 16,000 to 1500 Years before Present*. Advances in Anthropology, Vol.2, No.2, 87-105, doi:10.4236/aa.2012.22010, 2012.

[6] Alix Boc, Alpha Boubacar Diallo, Vladimir Makarenkov, *T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks*. Nucleic Acids Research (2012) 40 (W1): W573-W579. doi: 10.1093/nar/gks485, 2012.

[7] Wikipedia, *List of Y-STR markers*. Date accessed: 2014-08-15.

[8] Bruce Walsh, *Details on the various assumptions used in computing TMRCA*. Family Tree DNA, 2002, Date accessed: 2014-08-07.

[9] Bruce Walsh, *Estimating the Time to the Most Recent Common Ancestor for the Y chromosome or Mitochondrial DNA for a Pair of Individuals*. Genetics Society of America, 2001.