Phylofriend User Guide

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

March 12, 2016

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[6] 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[6] 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 github.com/yogischogi/phylofriend`

5. Install the program with
   `go install github.com/yogischogi/phylofriend`

3 Command Line Options

Command line options may be given in arbitrary order.

-help Prints available program options.

-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.

-labelcol Number of the column that is used for labels 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 and performs TMRCA calculation.

-phylipout Filename for the distance matrix that can be fed into the PHYLIP[6] program.

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

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

-nmarkers Uses only the given number of markers for calculations.

-gentime Generation time.

-cal Calibration factor.

-reduce Reduces the number of persons by the given factor (for large numbers of samples).

-statistics Prints marker statistics.

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 -mrin 67-average.txt`
   Here we use a set of average mutation rates for 67 markers. The file *67-average.txt* can be found in the directory *phylofriend/mutationrates/*. The result is a matrix that contains the number of generations as a measure for genetic distances between the persons.

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[3] web-server. You can copy the contents of the file *outtree* into the Trex window.

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 by using the *labelcol* option. Suppose your second column contains names. You can create a distance matrix with names instead of IDs by typing

```
phylofriend -personsin persons.csv -labelcol 2
-phylipout infile -mrin 67-average.txt
```

Due to compatibility issues with other programs the labels must be 10 characters long and may only contain 8-bit characters. Phylofriend will apply a transformation to make sure that the requirements are fulfilled but the result is sometimes a bit strange.

You can also use the *labelcol* 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
-mrin 67-average.txt
```

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

4.3 Use a Specific Set of Mutation Rates

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 [9]. If you like to compare on 67 markers or 111 markers you can use

```
phylofriend -personsin persons.csv -phylipout infile
-mrin 67-average.txt
```

or

```
phylofriend -personsin persons.csv -phylipout infile
-mrin 111-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 genetic distances in years.

Phylofriend provides two options for data calibration: *gentime*, the generation time 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.

A generation time of 32 years has proven to show good results [1]. You can use it by typing

```
phylofriend -personsin persons.csv -phylipout infile
-gentime 32 -mrin 67-average.txt
```

In reality an optimal result is often hard to archieve. Especially within the range of genealogical time frames (about 400 years) and only small numbers of persons who have tested, you are often left with a large statistical error. Even worse, the method of Y-STR counting has numerous pitfalls, some of which are discussed in [7].

If you have a reliable paper trail or a well defined historic event you can calibrate your data using the *cal* option. With *cal* you just provide an additional calibration factor that is multiplied to the calculated genetic distances, for example

```
phylofriend -personsin persons.csv -phylipout infile
-gentime 32 -mrin 67-average.txt -cal 1.2
```

multiplies all genetic distances by a factor of 1.2.

4.5 Count Mutations

If you want to count mutational differences between persons, for example on a 37 marker scale, you can archieve this by typing

```
phylofriend -personsin persons.csv -phylipout distancecount.txt
-nmarkers 37 -cal 37
```

The *nmarkers* option restricts all calculations to the given set of markers (usually 37, 67 or 111). If a person has not tested for enough markers, he is excluded from the calculation. Because Phylofriend uses average values internally, all results are devided by 37 in the previous example. To get whole numbers we must compensate for this effect. This is done by multiplying the results with 37 using the *cal* option.

The internal use of average values seems complicated at first but it allows Phylofriend to compare persons who have tested for different numbers of markers.

Phylofriend's results may also differ from other programs because Phylofriend uses it's own hybrid mutation model.

An alternative approach to mutation counting is using a set of mutation rates that was especially created for that purpose, for example:

```
phylofriend -personsin persons.csv -phylipout distancecount.txt
-mrin 37-count.txt
```

The difference to the previous example is that the *nmarkers* option in the previous example makes sure that all persons have tested for all 37 markers, while the use of the *37-count* mutation rate also accepts persons who have tested for a lower number of markers. The result is an estimate. This option was introduced because next generation sequencing reveals Y-STR results for about 500 markers. However due to the technical restrictions of this method, different persons usually get results for different, but largely overlapping, marker sets.

4.6 Marker Statistics

Phylofriend can give you detailed statistics about marker values. It calculates the minimal and maximal value for each marker and the number of occurrences. You can activate the statistical output by using the `-statistics=true` option, for example

```
phylofriend -personsin=s11481.csv -statistics=true -nmarkers=6
```

prints out the results for the first six markers (`-nmarkers=6`). In our example it is assumed that the persons' data is in a file called *s11481.csv*. This is the output:

```
DYS393, Min: 13, Max: 13, 13:15,
DYS390, Min: 24, Max: 24, 24:15,
DYS19, Min: 14, Max: 14, 14:15,
DYS391, Min: 10, Max: 11, 11:14, 10:1,
DYS385a, Min: 11, Max: 12, 11:10, 12:5,
DYS385b, Min: 12, Max: 14, 12:1, 14:14,
```

Each line contains the marker name and the minimal and maximal values. After that each marker value is listed together with it's frequency, separated by a colon.

Let's take DYS391 as an example. It has two values, 10 and 11. The value 10 occurs 1 time and the value 11 occurs 14 times.

4.7 Use Data from YFull

YFull usually reports about 400 Y-STR marker results. To use these results put the YFull result files for every single person into a directory, for example *inputdir*. Call Phylofriend by typing:

```
phylofriend -personsin inputdir -phylipout infile -modal
-gentime 32 -mrin 500-average.txt
```

This command will use up to 500 markers for comparison. The mutation rates for 500 markers are still experimental.

4.8 Extract Data from a Spreadsheet in CSV format

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 you can use the *labelcol* option:

```
phylofriend -personsin persons.csv -txtout persons.txt
-labelcol 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 *nmarkers*. 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
-nmarkers 111
```

5 Technicalities

5.1 Source Code Documentation

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

- http://godoc.org/github.com/yogischogi/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[13]. Phylofriend uses a hybrid mutation model. Most markers are calculated using the stepwise model, palindromic markers are calculated as described in [4].

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.

Average Mutation Rates

- 12-average.txt
- 37-average.txt
- 67-average.txt
- 111-average.txt
- 390-average.txt (experimental)
- 500-average.txt (experimental)

In these files average mutation rates are used for the corresponding set of markers. The mutation rates for 12, 37, 67 and 111 markers were taken from [9].

The mutation rates for 390 and 500 markers are experimental. They were derived by comparing YFull data of the R1b-M343 (xP312 xU106) project to Family Tree DNA data from members of the R1b-M269 (P312- U106-) project. 500-average.txt uses all known Y-STR markers that are reported by YFull and Family Tree DNA. 390-average.txt leaves out most of the palindromic markers. This marker set should suffer less from saturation effects.

These mutation rates are useful for genealogical purposes. During genealogical time frames (about 400 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 appropriate weight on very stable markers. The mutation rates were taken from Wikipedia[12]. 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.

Mutation Rates for Marker Counting

- 12-count.txt
- 37-count.txt
- 67-count.txt
- 111-count.txt
- 390-count.txt
- 500-count.txt

These mutation rates can be used for marker counting between different persons, even if they have not been tested on the exact same markers. Phylofriend will give you an estimate about the expected marker difference on the specified scale.

This approach is recommended for large marker sets from next generation sequencing.

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 YFull Format

YFull files have a name like *STR_for_YF01234_20160222.csv*. Each line of the file contains the result for a single marker and sometimes additional information. Example:

```
DYS390;24;
DYS391;11;
DYS392;14;?
DYS393;13;
```

Although the file is in CSV format, semicolons are used instead of commas as separators. This may cause trouble if you try to add additional results by hand. Phylofriend does not care if a marker is provided multiple times. The last occurrence of a name is considered the valid one. Thus you can add results from other testing companies just by adding them to the end of the file.

Phylofriend tries to extract a person's ID from the filename. If a file is named *STR_for_YF01234_20160222.csv*, the ID will be *YF01234*.

If you want to provide your own files, you do not need to stick to the YFull naming convention. Just use the desired ID as a filename like *ID1234.csv*, but the file must end in *.csv*.

5.7 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

Parts of the human Y chromosome are only inherited from father to son [1, 5]. Usually these parts 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.

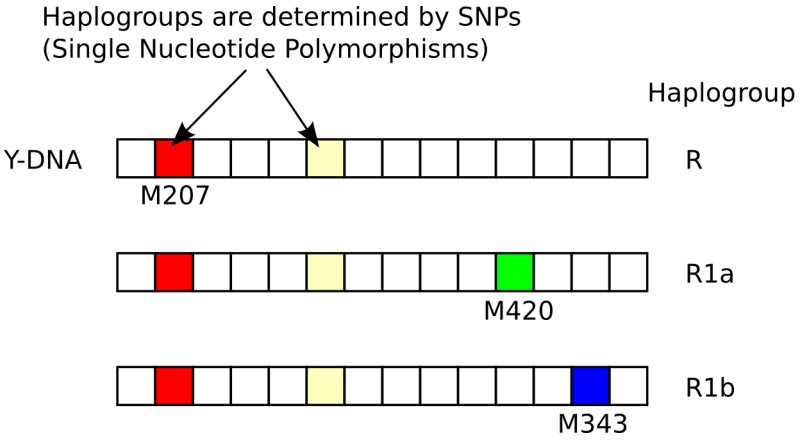

Figure 1: A set of single mutations defines a person's haplogroup. Persons who belong to the same haplogroup usually share a common ancestor within a few thousand years.

Figure 1 illustrates a typical situation. Many Europeans belong to haplogroup R which is characterized by the mutation M207. 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 occur approximately once in a hundred years [1] but usually only SNPs that are older than 1000 years are listed at ISOGG [10]. Under normal circumstances Persons who belong to the same haplogroup 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). Newer tests like Family Tree DNA's Big Y [5] also make it possible to find relatives within a genealogical time frame.

6.2 Haplotypes

Most people want to know more about family relationships. Relationships are often verified by STR (Short Tandem Repeats) mutations. STRs are repetitions of certain genetic patterns. They group people into haplotypes.

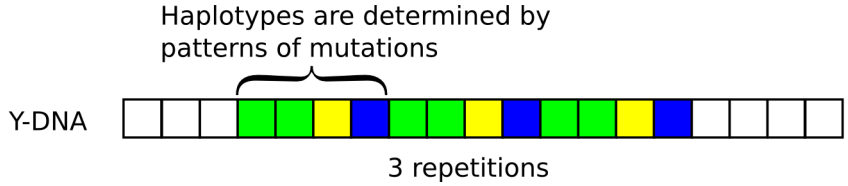

Figure 2: A person's haplotype is defined by patterns of mutations. Persons with the same haplotype usually share a common ancestor within a few hundred years if they take a standard test with 37 markers or more.

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 of father and sons. Because we know the birth dates we would put the father

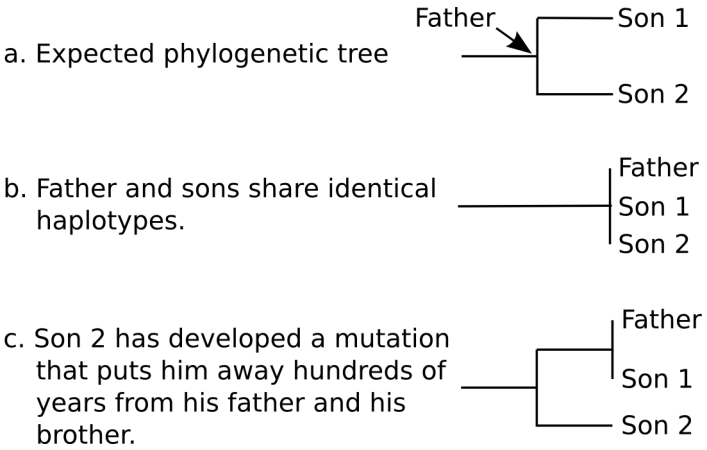

Figure 3: Phylogenetic trees group persons together according to their genetic distances. Genetic distances are often associated with time scales but this is only true for long time spans.

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 [9] 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 Distance

There are many ways to calculate genetic distances. Here we use the method of mutation counting as described by Anatole Klyosov in [8]. 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. But mutations occur by coincidence. Thus for a precise measurement we need large numbers of mutations. Different family lineages often expose different mutation rates. So the standard mutation rates should be considered as a first estimate. You can use Phylofriend's *cal* option to adjust your data to historic events.

Now we can use equation 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 [8]. 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 [8].

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[14]. 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.

The whole story is of course more complicated than the simple calculation presented here. Different mutation models exist and some genetic markers should be treated in a special way. Phylofriend uses a hybrid mutation model that is a mixture of the stepwise mutation model and the infinite alleles model. Both models are explained by Bruce Walsh in [13].

6.5 Modal Haplotype

If we have a group of people we are interested in the haplotype of their most recent common ancestor. This is called the modal haplotype, base haplotype or ancestral haplotype. The following table shows the haplotypes of three persons on the 6-marker scale and their modal haplotype.

	DYS393	DYS390	DYS19	DYS391	DYS385a	DYS385b
modal	13	24	14	11	11	14
Peter	13	24	**15**	11	11	14
Carl	13	24	14	11	**12**	14
Clas	13	**23**	14	**12**	11	14

The modal haplotype can easily be calculated. If we use equation 4 to calculate the genetic distance in years for just a single marker we see that it takes 12500 years on average for a single marker to mutate only once. Thus for most practical purposes a single marker rarely mutates at all. So for a single marker we just look which value occurs most often and take this as the modal value.

Phylofriend uses this simplified approach. For many thousands of years it is no longer valid. Calculating the median values would be a better choice then. The problem with the median approach is that some markers show rather big leaps when mutating. This could lead to wrong results. I recommend that you always take a good look at your data to see which method would be appropriate.

6.6 Age Calculation

6.6.1 Mutation Counting

The simplest method to calculate the time to the most recent common ancestor (TMRCA) for a group of people is mutation counting as described in great detail by Anatole Klyosov in [8]. Here a give a short overview only.

First we calculate the modal (ancestral) haplotype. Then we count the mutations of every single group member relative to the modal haplotype and

take the average. This should be a good estimate for the group's genetic distance to their ancestor.

Again we use the definition of the mutationrate 1 but this time we use the average value:

$$k = \frac{\bar{m}}{gy} \tag{6}$$

k: Mutation rate per marker and generation
$\bar{m}$: Average number of mutations relative to modal haplotype
g: Number of generations
y: Number of Y-STR values per haplotype

As mutations occur by coincidence equation 6 is only valid for sufficiently high values of $\bar{m}$. By using the definition of the average, equation 6 is equivalent to

$$k = \frac{\sum_{i=1}^{N} m_i}{Ngy} \tag{7}$$

or

$$g = \frac{\sum_{i=1}^{N} m_i}{Nky} \tag{8}$$

g: Number of generations
m_i: Number of mutations of haplotype i relative to modal haplotype
N: Number of haplotypes (persons)
k: Mutation rate per marker and generation
y: Number of Y-STR values per haplotype

Equation 8 is the time to the most recent common ancestor in generations. If we know the generation time d we can write this (see equation 3) as

$$t = d\frac{\sum_{i=1}^{N} m_i}{Nky} \tag{9}$$

t: Time to most recent common ancestor (TMRCA)
d: Generation time
m_i: Number of mutations of haplotype i relative to modal haplotype
N: Number of haplotypes (persons)
k: Mutation rate per marker and generation
y: Number of Y-STR values per haplotype

The following example shows how to use the formula but it is intended as an easy to understand educational exercise. For real world applications more haplotypes (rule of thumb: use at least 5) and a bigger number of average mutations (4 or more is quite good) is needed.

The table on page 21 contains three haplotypes on the 6-marker scale ($N = 3$, $y = 6$). Peter and Carl differ by one mutation to the modal haplotype and Clas by two. We use a generation distance of 25 years ($d = 25$) and the average mutation rate on the 6-marker scale is 0.002[8]. So we get

$$t = 25 \cdot \frac{1 + 1 + 2}{3 \cdot 0.002 \cdot 6} = 2778 \; years \tag{10}$$

Of course the simple method of mutation counting comes not without problems. The biggest one is the correct distribution of the population sample. For example if you want to calculate the TMRCA for a family which is just a few hundred years old and add a person by accident who is at a distance of several thousand years your TMRCA value will be much too large.

On the other hand if you want to calculate the age of a whole haplogroup and add a large number of members from one family who are all very close together your TMRCA value will be much too small. So this method must be handled with care.

A more general problem is that mutation rates are derived from population samples and there is no guarantee that they fit to your data set. So whenever possible try to find some historical event to calibrate your data.

6.7 From Distances to Trees

Phylofriend computes genetic distances between different persons. These distances can be used to derive phylogenetic trees using standard algorithms that are implemented by the PHYLIP[6] program.

Genetic distances are often represented by two-dimensional drawings that are easy to understand. Here we label the distance components as *Component 1* and *Component 2*. Usually the values for each component are derived from many markers by using principal component analysis[11].

Figure 4 shows the genetic positions of several persons and the corresponding phylogenetic trees. In the first example Carl and Clas are close together. So they are also grouped together in the tree. Peter is at a longer distance but the distance to Carl and Clas is the same. Thus he is put into a separate branch. The genetic distance between the persons is represented by the length of the horizontal lines in the tree.

The second example is slightly more complicated. Carl and Clas are still close together but Peter is at a different distance to each of them. The trees

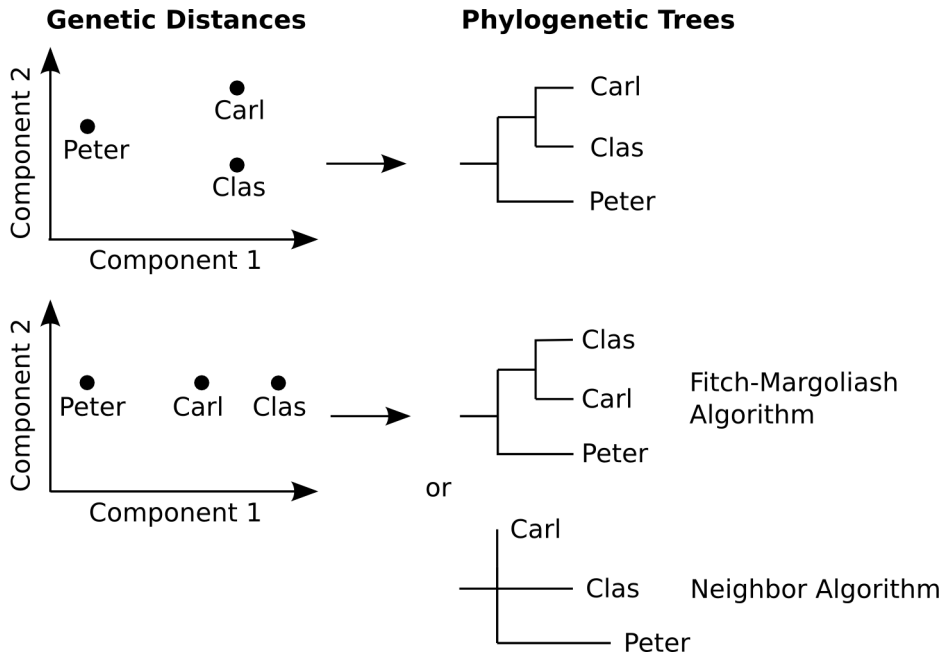

Figure 4: Genetic distances can be represented by phylogenetic trees. Different algorithms often yield different results.

were constructed by using two different methods, Fitch-Margoliash (kitch program) and Neighbor. Fitch-Margoliash groups Carl and Clas together and puts Peter at a medium distance from both of them. This is a sensible approach because some people develop many mutation and others develop less. So the ancestor of Carl and Clas was probably somewere in-between. However the resulting tree does not represent the genetic distances correctly but it displays a very nice representation of temporal evolution.

The neighbor method tries to represent the genetic distances as good as possible. In this case it is accurate but it gives you a bad idea of temporal evolution.

For a high quality data set both methods should yield similar results in grouping people together. The real problem is that mutations occur by coincidence and we often have a very sparse collection of descendants who are at large genetic distances to each other.

Figure 5 shows a more realistic situation. It represents the genetic positions of members from two different families. All members of one family descent from one common ancestor. Due to the laws of statistics most family members are close to his haplotype. So it can be calculated. It is called the ancestral or modal haplotype.

Other family member are at a greater distance from the ancestor. Some-

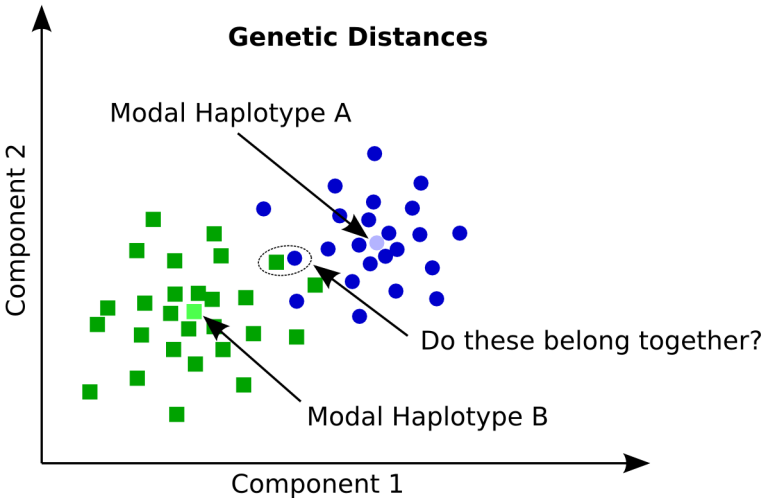

Figure 5: Haplotypes and genetic distances of two families. Although both lineages have a different modal (ancestral) haplotype the haplotypes of individuals from both families are sometimes close together. In such cases phylogenetic tree algorithms yield wrong results.

times members from different families come close to each other. In such cases all algorithms will give you wrong results.

This is the reason why it is so important to test for haplogroups before creating a phylogenetic tree. Haplogroups are caused by very stable mutations. Haplotypes often overlap. So it should be ensured that all persons in a tree belong to the same haplogroup.

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] 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.

[3] 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.

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

[5] Family Tree DNA, *Big Y White Paper*. Family Tree DNA, August 2014.

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

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

[8] 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.

[9] 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.

[10] ISOGG, *Listing Criteria for SNP Inclusion into the ISOGG Y-DNA Haplogroup Tree - 2015*. Date visited: 2015-10-06.

[11] Jonathon Shlens, *A Tutorial on Principal Component Analysis*. Center for Neural Science, New York University, New York City, Systems Neurobiology Laboratory, Salk Insitute for Biological Studies, La Jolla, 2009.

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

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

[14] 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.