PhyloGeoTool
Installation Manual

EWOUT VANDEN EYNDEN, PIETER LIBIN, KRISTOF THEYS,
GUY BAELE
Rega Institute for Medical Research, KU Leuven
May 2017

Contents

1 Installation

1.1 Supported platforms

PhyloGeoTool is supported on GNU/Linux, MacOS and Microsoft Windows. These installation instructions are tested on all of these platforms.

1.2 Prerequisites

Java

Install the Java Development Kit (JDK) with version $\geq$ 1.7.

For example, on Ubuntu Linux, Oracle Java 8 can be installed as follows:

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java8-installer
```

Tomcat

Install Tomcat version $\geq$ 7.

For example, on Ubuntu Linux, Tomcat 7 can be installed as follows:

```
sudo apt-get update
sudo apt-get install tomcat7
```

R runtime

Install the R runtime (version $\geq$ 3.2.1).

For example, on Ubuntu Linux, the R runtime can be installed as follows:

```
sudo apt-get install r-base r-base-dev
```

PPlacer and its dependencies

Install the python package manager **pip**.

For example, on Ubuntu Linux, **pip** can be installed as follows:

```
sudo apt-get install python-pip
```

Download and extract **BLAST v2.2.14** from `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/legacy/2.2.14/`. The location to the bin folder within the BLAST directory needs to be added to the global configuration file.

Install **taxtastic** as explained at `http://fhcrc.github.io/taxtastic/installation.html`.

Install **Mafft v7.271** to align sequences to be classified by PPlacer. For example, on Ubuntu Linux, Mafft can be installed as follows:

```
sudo apt-get install mafft
```

Install PPlacer v1.1 (website: `http://github.com/matsen/pplacer/releases`).

Phylogenetic tree

A rooted phylogenetic tree in either Nexus or Newick format. It is important that the tree is **binary**.

CSV file

A comma-separated value (CSV) file, containing an ID column that contains identifiers that correspond to the taxa names of the provided phylogenetic tree. Remaining columns may contain additional information/annotation for the taxa in the phylogenetic tree. Examples of such annotations are: geographic location, virus genotype/subtype and patient attributes (age, ethnicity, ...).
Information on the geography of taxa can be visualized in a map (Google Charts). For this to work, the column containing the geographic information needs to be configured with the "visualizeGeography" property in the XML configuration file (for more information, see section 1.4). The geographic values need to be formatted with the code or country name as defined in the ISO_3166-1_alpha-2 standard [1].

1.3 Create a distance matrix and determine clusters

A major goal of the PhyloGeoTool is to cluster the tree and to index the taxa annotations, of which the computation takes a long time and the runtime depends on the size of the phylogenetic tree (computational complexity is $\mathcal{O}(n^3)$). However, this computation can be performed before the installation of the web-tool (using PreRender.jar).

[1] `https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2`

To be able to support this computation, a distance matrix needs to be constructed (using DistanceMatrix.jar). Performing this clustering procedure before the installation of the web-tool allows the web-tool to render both clusters and annotations instantaneously. To infer a distance matrix based from the phylogenetic tree, the program DistanceMatrix.jar is used. DistanceMatrix.jar can be downloaded from our website [2] or built from source (see section 2). The following command uses DistanceMatrix.jar to generate a distance matrix (to be stored in a file called distances.csv in the command below) from a previously constructed phylogenetic tree in Newick format (i.e. tree.newick):

```
java -jar DistanceMatrix.jar tree.newick distances.csv
```

To subsequently partition the tree in clusters, the program PreRender.jar is used. PreRender.jar can be downloaded from our website or built from source (see section 2). PreRender.jar accepts the following parameters:

- tree.newick: location of the phylogenetic tree used in the PhyloGeoTool.
- attributes.csv: location of the CSV file that connects nodes in the tool to attributes [3].
- distances.csv: location of the distance matrix that was generated from this tree.
- /path/to/cluster_output: location of the folder where this the PreRender.jar program has to write its output files. The files are stored in four different folders.
    1. clusters: contains the files that stores the best clustering representation in XML format for a specific level.
    2. r: contains the graphs (SDR, First Derivative and Second Derivative) which are generated during the clustering phase of the tool.
    3. treeview: contains images that display a colored phylogenetic tree, where each color is parallel to the cluster shown on that level.
    4. xml: contains a summary of the attributes which are displayed at each specific level of the tool.
- /path/to/rBinary: link to the exact location of the R executable.

[2] `https://github.com/rega-cev/phylogeotool/releases/tag/v1.0.0`

[3] The identifiers in the CSV file (in the ID column) have to correspond with the identifiers of the nodes in the tree.

- /path/to/folder_rScripts: link to the folder which contain SDR.R, FirstDerivative.R, sgolay.R and SecondDerivative.R (i.e. https://github.com/rega-cev/phylogeotool/tree/master/scripts).

PreRender.jar will create the necessary directories in the folder with name 'folder_output'. If those directories already exist, a warning is issued to the user and the process will be aborted. For example, the following command uses PreRender.jar to perform all the necessary clustering steps (which can be time-consuming) such that the PhyloGeoTool doesn't have to perform these at run time:

```
java -jar PreRender.jar phylogenetic.tree csvFile
distance_matrix folder_output rBinary folder_rScripts
```

1.4 Configure PhyloGeoTool

In the /etc folder, create a directory 'phylogeotool', which is the default location for the configuration file.
In the case of a Windows installation, create this directory under C:\\Program files such that the full path becomes C:\\Program files \phylogeotool \.
You can also set your own preferred folder by extending the web.xml file in your deployed application with the following information.

```
<context-param>
    <param-name>conf-dir</param-name>
    <param-value>/path/to/conf/file/phylogeotool</param-value>
</context-param>
```

The param-name should be 'conf-dir' and the param-value indicates where the program has to look for the configuration. An example of such a configuration file can be found here: `https://github.com/rega-cev/phylogeotool/blob/master/examples/global-conf.xml`. In this example file, the various configuration concepts are explained. In the XML file, we assume a user name 'phylogeo' to perform the installation. Edit the example global-conf.xml file to correctly set up all the necessary paths.

1.5 Install the PhyloGeoTool WAR file in Tomcat

The PhyloGeoTool WAR file (i.e. phylogeotool.war) can be downloaded from our website [4] or built from source (see section 2).

Copy the phylogeotool.war to the webapps folder of Tomcat and start (or restart) Tomcat:

[4] `https://github.com/rega-cev/phylogeotool/releases/tag/v1.0.0`

```
sudo cp phylogeotool.war /var/lib/tomcat7/webapps/
cd /var/lib/tomcat7/
sudo service tomcat7 restart
```

This enables browser access to a localhost version of the PhyloGeoTool. Open a browser and enter the following URL: `http://localhost:8080/phylogeotool/PhyloGeoTool`. The browser should show something similar to Figure 1.

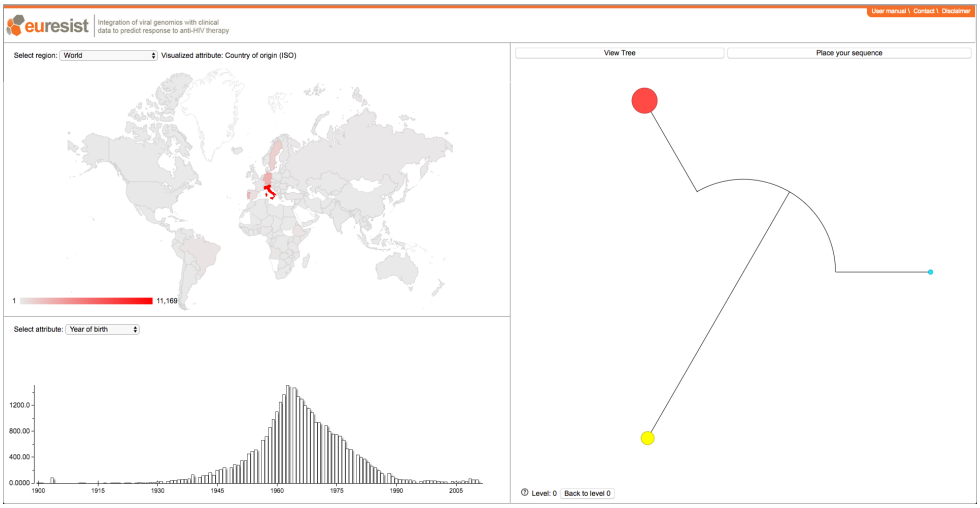

Figure 1: Screenshot of the PhyloGeoTool application when the application is started for the first time.

The PhyloGeoTool can be installed in a Tomcat container. When the application is to be used on a local computer, the container can safely be accessed with a local web browser. However, to host your own instance to be accessed by external users (i.e. an online instance), it is important to setup your instance in a secure way. It **is insecure** to let Tomcat run directly on **port 80**. Therefore, you need to run it on a port that does not require root access (e.g. port 8080) and use an Apache web-server as a reverse proxy as explained at [5].

2 Building from source

2.1 Prerequisites

Install the Java Development Kit (JDK) with a version $\geq$ 1.7.

For example, on Ubuntu Linux, Oracle Java 8 can be installed as follows:

[5] `https://wiki.apache.org/httpd/TomcatReverseProxy`

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java8-installer
```

Git

Install git.

For example, on Ubuntu Linux, git can be installed as follows:

```
sudo apt-get install git-all
```

Ant

Install Ant version $\geq$ 1.9.4 as we will use it to build the source code.

For example, on Ubuntu Linux, Ant can be installed as follows:

```
sudo apt-get install ant
```

2.2 Clone the git repository and build the JAR files and WAR file

We here outline the various steps necessary for a successful build of the project.

- Clone the git repository:

  ```
  git clone https://github.com/rega-cev/phylogeotool/
  cd phylogeotool
  ```

- Start the build process by running ant

  ```
  ant
  ```

- This build process generates a phylogeotool.war, DistanceMatrix.jar and PreRender.jar in the dist directory.

3 Contact information and support

The PhyloGeoTool is developed and maintained by the Rega Institute KU Leuven. You can contact us on phylogeotool@kuleuven.be