

Phylogena Manual

1. SYSTEM REQUIREMENTS
2. FAST INSTALLATION (WINDOWS ONLY)
3. EDIT THE STARTUP FILE
 - *Windows*
 - *LINUX*
4. USAGE BASICS
 - *Importing sequences*
 - *Start an analysis*
 - *Viewing results*
5. SELECTION RULES
6. SELECTING ALIGNMENT SOFTWARE TO USE
 - *ClustalW*
 - *Kalign*
 - *MUSCLE*
 - *Mafft*
 - *POA*
 - *T-Coffee*
7. SELECTING TREE RECONSTRUCTION SOFTWARE TO USE
 - *PHYLIP*
 - *Quicktree*
 - *PhyML*
8. INSTALLING A DATABASE
9. INSTALLATION OF PHYLOGENA
 - *Installing the "Easy" version*
 - *Install third-party executables*
 - *Install Phylogena*
 - *Install BLAST databases*
 - *Edit the config file*
 - *Run biojava indexing for each database*
10. TROUBLESHOOTING

Last modified: 20/06/2007, Bánk Beszteri

Phylogena is a software pipeline that performs similarity searches against a local copy of SWISSPROT (or another sequence database) using BLAST, selects hits based on their quality and (taxonomic and functional) diversity, creates a multiple alignment from the selection (using ClustalW or another multiple alignment program) and calculates a phylogenetic tree (using PHYLIP, Quicktree or PhyML) from that. The primary aim of the pipeline is to facilitate functional annotation of unknown ORFs using a phylogenetic approach. It was written by Kris Hanekamp for his master's thesis under the supervision of Klaus Valentin, Uta Bohnebeck and Otthein Herzog. It has since gone through improvements by Christophe Garnier and me (bb).

Phylogena uses the following third-party components – see the web addresses listed to obtain them or further information about them:

Java libraries (you don't need to get these, they are included with PhyloGena):

- 2prolog (<http://www.alice.unibo.it:8080/tuProlog/>) # gpl
- biojava (<http://biojava.org>) # gpl
- jalview (<http://www.jalview.org/>) # gpl
- ATV (<http://www.genetics.wustl.edu/eddy/forester/>) # 'as is'?

Other third-party executables (you need to download them unless you're taking our "Easy" version for Windows, which includes these programs as well):

- NCBI Blast (<ftp://ftp.ncbi.nih.gov/blast/executables/LATEST>)
- Alignment programs: ClustalW, Kalign, Mafft, or MUSCLE (see below)
- Phylogenetic analysis programs: PHYLIP, Quicktree or PhyML (see below)

1. System requirements

The software has mostly been tested on Windows 2000 and XP and we have also got it running on different LINUX systems. We do not use Macs, and have made little effort to get it running on this platform as yet. To get started on Windows, we recommend you take the "Easy" version. We would be happy to hear from you if you tried (or succeeded) to run it on a Mac.

We recommend that you use a relatively recent computer (> 1GHz CPU, > 256 K RAM) to be able to analyse medium sized (10-100 sequences) datasets within a reasonable time. An installation with SWISSPROT needs about 1 GB of disk space.

Phylogena requires a java runtime environment, version 1.4 or 1.5. Java is installed on most computers these days (on Windows, check under Program Files\Java); if you haven't got it yet, you can obtain it from <http://java.sun.com> (look for Java Runtime Environment under Downloads).

2. Fast installation (Windows only)

For a fast installation of Phylogena on Windows including SWISSPROT, NCBI BLAST, ClustalW and PHYLIP, download phylogenaEasy.zip, and unzip it under C:\. Edit your start-up file (see below), and you are ready to go.

3. Edit the startup file

➤ Windows

The startup file for Windows is called `phylogena.bat`. You have to specify the path to the java virtual machine (the directory, in which your java executables – under Windows, `java.exe` – can be found) as `LOCAL_JAVA` in this file, like

```
set LOCAL_JAVA="C:\Program Files\Java\j2re1.4.2_10\bin\java"
```

The quotes are only needed if there is a space in one of the path, like `Program Files`.

After this, you should be able to run Phylogena by double-clicking on `phylogena.bat`.

➤ LINUX

We provide an example startup file for LINUX (`phylogena.sh`). You might need to adjust the path to java to your system (`/usr/bin/java` in the example file).

4. Usage basics

First, choose which DB you want to use under Extra / Choose database. The “Easy” version comes with a single database (SWISSPROT), so this is only a concern for you if you have set up (a) custom database(s) – see below for information on how this can be done.

➤ Importing sequences

Then you can import FASTA formatted sequences by clicking on File / Import fasta sequences. You are then being asked whether the sequences to be loaded are nucleotide or protein; choose and click OK.

➤ **Start an analysis**

To start an analysis, you can either click on Rules / Analyse whole project, or mark a sequence and click on Rules / Analyse a single query.

You will first have to decide whether you want to run simple BLAST to collect similar sequences or you want to perform recursive BLAST (re-blast hits from a first round of BLAST; this might produce more hits and can be useful in some cases where a simple blast returns too few results).

Then you can choose the rule to select a subset of BLAST hits to be used for multiple alignments and phylogenetic analyses (see below for a short description of them).

At last, you can specify a number of parameters which are in part general (e.g., name of analysis, e-value limit, alignment and tree reconstruction program to be used, number of bootstrap replicates etc.), in part specific for the selection rules. If you then click OK, the analysis starts.

➤ **Viewing results**

If an analysis has completed, you can look at the alignments (View / Show alignment) and trees (View / Show tree) produced after selecting a query. You can check the BLAST results and modify the selections made by the program by navigating in the project tree on the left hand side to Result set. If you modified the sequence selection, you can calculate a new tree by first clicking on Analyse / Calculate alignment and then on Analyse / Calculate NJ tree (you have the choice between NJ, UPGMA and ML). If you have prepared more than one alignment for a query, you have to select the one you want to calculate a tree from. Similarly, if you have calculated multiple trees for a query and want to display one of them, select the tree you want to see on the left hand side of the screen instead of only selecting the query name.

5. Selection rules

The following selection rules are currently implemented in the pipeline:

- choosing the best x hits (based on e-value);
- choosing the best x hits from each taxon at a user-defined taxonomic level;
- choosing the best x hits from each taxon at the lowest taxonomic level where the number of selected hits does not exceed a user-defined limit;
- intelligent branching.

The combined selection rule referred to as intelligent branching uses different combinations of selection rules depending on the composition of the BLAST result. It distinguishes four basic cases. In the case of a “normal” BLAST result (i.e., not categorized into any of the other three cases), diversity selection concerning functional annotation and taxonomy is followed by a quality based selection. In the

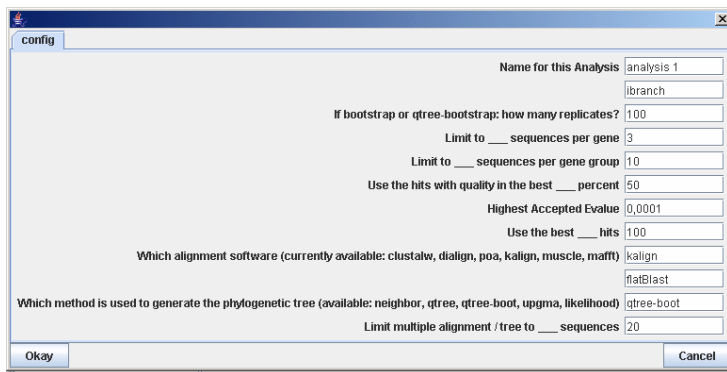
following special cases, this basic selection is complemented by further filtering steps:

- if there are many (> 100) BLAST hits, the above steps are preceded by a first rough quality selection, removing the worst hits;
- if the quality distribution of the BLAST result is bi- or multimodal, the hits in the lower quality spectrum are removed before the first diversity selection;
- if the BLAST result is very heterogeneous concerning both gene function and e-value, the “normal” selection is preceded by filtering out sequences which have functional annotations not occurring among the best hits.

6. Selecting alignment software to use

PhyloGena is written in a modular fashion, so that plugging in different multiple alignment and tree reconstruction programs is made relatively easy. Below is a list of which programs are interfaced from PhyloGena at the moment. In order to be able to use any of them in your local installation, you need to perform the following steps:

1. install the program on your computer (see installation instructions at their respective source websites / manuals)
2. specify their location in your phylogena config file (phylogena.config by default)
3. type the name of the program in the respective field of the third dialog when starting an analysis, which could look like:



➤ ClustalW

ClustalW can be obtained from

<http://www.cf.ac.uk/biosi/research/biosoft/Downloads/clustalw.html>.

When installed on your computer, specify the path to the executable as

```
clustalw.executable
```

in the config file. Type “clustalw” into the field for choosing alignment program in the config window when starting an analysis.

➤ **Kalign**

Kalign can be obtained under <http://msa.cgb.ki.se/cgi-bin/msa.cgi>. The last time we checked there was only source code available, but it is easily compiled also on Windows under cygwin. (Note that when using a cygwin-compiled version, you'll need to have the file cygwin1.dll in your path.)

Once you installed Kalign, you need to specify where it is in your config file as

```
kalign.executable
```

and type "kalign" into the field for choosing alignment program in the config window when starting an analysis.

➤ **MUSCLE**

MUSCLE executables for Windows as well as UNIX can be obtained from <http://www.drive5.com/muscle/>. If you have it, you need to specify its location in your config file as

```
kalign.executable
```

and type "muscle" into the field for choosing alignment program in the config window when starting an analysis.

➤ **Mafft**

Mafft can be obtained at <http://align.bmr.kyushu-u.ac.jp/mafft/software/>. Mafft is a genuine UNIX program, i.e., the command line front-end is a shell script, making it a bit complicated to get running on Windows. The mafft webpage gives installation instructions for Windows using Cygwin. Our solution in PhyloGena is to use such an installation and run mafft using cygwin bash. For this solution to work, you need to have not only cygwin1.dll in your path, but also cygwin bash installed on your system. You will also need to specify the path to the latter in the configuration file, and the cygwin path to your mafft installation, like:

```
mafft.executable = /usr/local/bin/mafft  
cygwin.bash = C:/cygwin/bin/bash
```

On LINUX, you only need to specify the mafft path. Once you have set it up, you can type "mafft" into the field for choosing alignment program in the config window when starting an analysis.

➤ **POA**

POA can be obtained from

<http://www.bioinformatics.ucla.edu/poa>.

Once installed, specify the path to your executable as

poa.executable

Type “poa” into the field for choosing alignment program in the config window when starting an analysis.

➤ **T-Coffee**

T-Coffee can be obtained from

http://www.igs.cnrs-mrs.fr/Tcoffee/cnotred/public_html/Projects_home_page/t_coffee_home_page.html.

The config file keyword is

tcoffee.executable

Type “tcoffee” into the field for choosing alignment program in the config window when starting an analysis.

7. Selecting tree reconstruction software to use

In order to be able to use any of them in your local installation, you need to perform the same steps as for using a multiple alignment program:

1. install the program on your computer (see installation instructions at their respective source websites / manuals)
2. specify their location in your phylogena config file (phylogena.config by default)
3. type the name of the program in the respective field of the third dialog when starting an analysis (see above figure)

When running bootstrap analyses, you can also specify the number of bootstrap replicates to be run in the same dialog window.

➤ **PHYLIP**

You can use the PHYLIP package in PhyloGena to calculate UPGMA, neighbor joining (NJ) and maximum likelihood (ML) trees and to bootstrap them. However, especially if you want to run bootstrap analyses, we recommend that you use Quicktree for NJ and PhyML for ML as they are considerably faster. In any case, this is what you need to do in order to use PHYLIP. First, install the PHYLIP package (it can be obtained from <http://evolution.genetics.washington.edu/phylip.html>). Then specify the location of the PHYLIP binaries folder on your system in the phylogena config file as

phylip.directory

For choosing one of the PHYLIP programs for an analysis, type “upgma” for selecting PHYLIP UPGMA, “neighbor” for selecting PHYLIP NJ, “bootstrap” for PHYLIP NJ bootstrap, “likelihood” for PHYLIP ML, and “bootstrapML” for selecting PHYLIP ML bootstrap as a tree reconstruction method in the last config window when starting an analysis.

➤ **Quicktree**

Quicktree is a program for fast calculation of neighbor joining trees (and for bootstrapping). You can obtain it from <http://www.sanger.ac.uk/Software/analysis/quicktree/>. Once you’ve installed it, specify the path to where the executable lives in your phylogena config file as

```
qtree.executable
```

To use it for an analysis in PhyloGena, type “qtree” (for simple analysis) or “qtree-bootstrap” (for bootstrapping) in the last config window when starting an analysis.

➤ **PhyML**

Because maximum likelihood analyses with PHYLIP are unaffordably time-consuming, we also implemented an interface to the much faster ML program PhyML. PhyML (executables for Windows as well as LINUX) can be obtained under <http://atgc.lirmm.fr/phyml/>. The path to the executable should be specified in the phylogena config file as

```
phyml.executable
```

There are several command line parameters you can set for running PhyML. To set these parameters, there are two more fields in the phylogena config file, called `phyml.params1` and `phyml.params2`, where you can simply include the command line switches of your choice for PhyML (in the same form as you would specify them on the command line; with the exception that the parameters before and after the number of bootstrap replicates [which is in position 4] are split into two config entries). To see what your options are, check out the documentation on the Phyml web site (under User’s guide, go to the “command line interface”).

8. Installing a database

You can install other databases instead of / besides SWISSPROT. Phylogena’s advanced sequence selection capabilities (intelligent branching) require the use of SWISSPROT (as this rule uses the SWISSPROT gene names for grouping hits), and taxonomy-based selection rules require EMBL formatted sequence flat files for the moment (e.g. SWISSPROT plus TrEMBL, available for download from EMBL). Simple fasta databases can also be installed, but you will only be able to use the rule selecting the x best BLAST hits with them (a feature we’re planning to implement is to be able to attach taxonomic information also to simple fasta sequence files). Installation of a new sequence database is described in the following sections, under Installing BLAST databases.

9. Installation of PhyloGena

There are currently two ways to install PhyloGena. On Windows, you can take the “Easy” version, which comes together with all executables and a copy of SWISSPROT and is configured so that it can be run almost “out of the box”. On LINUX, you need to go the hard way and install software and databases for phylogena yourself as described below.

➤ Installing the “Easy” version

To install Phylogena on a Windows system, the easiest possibility is to unpack phylogenaEasy.zip under C:.

A folder called phylogena will appear under C: with all the files required for PhyloGena, in a working setup. The only customization necessary is to specify the path to java.exe on your system in phylogena.bat (right click / edit). Replace the path in the following line with the path on your system, which could look like:

```
set LOCAL_JAVA="C:\Program Files\Java\jre1.5.0_06\bin\java"
```

Java is normally installed under Program Files\Java in a folder with a version name (like jre1.5.0_06). Java.exe can be found in the "bin" subfolder, i.e. most probably you only have to replace the version number in the above line (after \java and before \bin\java).

After this, you should be able to start phylogena by double clicking on phylogena.bat (a Windows command window will appear; do not close this as long as phylogena is running).

➤ Install third-party executables

Phylogena needs the following executables:

1. ncbi blast (<ftp://ftp.ncbi.nih.gov/blast/executables/LATEST>)
2. clustalw (<http://www.cf.ac.uk/biosi/research/biosoft/Downloads/clustalw.html> – on Windows: if the DOS version doesn't work for you, take the XP)
3. the PHYLLIP package
(<http://evolution.genetics.washington.edu/phyllip/getme.html>)
4. java virtual machine, version 1.4 or 1.5 (go to <http://java.sun.com/downloads> and choose Java Virtual Machine in case it is not yet installed on your computer)

Install them according to the instructions you find on the above websites – this simply means unpack them into a suitable directory, like C:/phylogena/bin/ (for 1-3) or run the installation program (for java).

Note / remember where you installed these programs, because you will have to specify their location to Phylogena in the config file (see below). Path to the java virtual machine has to be specified in the phylogena.bat file (also see below).

➤ **Install Phylogena**

This includes nothing more than unpacking the distribution file phylogena.zip into the location where Phylogena is supposed to run (e.g., into C:\phylogena). This contains the following files and directories:

1. three plain text files: `phylogena.bat` (to start the application on Windows), `phylogena.config` (which you'll have to edit to specify locations of your databases and third-party executables, see below) and `NCBI_BlastOutput.dtd`, which Phylogena needs for reading BLAST outputs, and
2. three directories, called `lib`, `src_pl` and `dtd`.

➤ **Install BLAST databases**

Phylogena's sequence selection features currently require a local copy of a SWISSPROT formatted flatfile sequence database. You have two choices concerning structuring your data. If you download swissprot / trembl especially for Phylogena, follow the instructions under 2 i). If you already have some components required by Phylogena in a central location (like fasta files or the blast files prepared with formatdb), go to point 2 ii).

a) Installing a SWISSPROT / EMBL type DB

First, decide where you want to keep the databases. A complete DB installation of SWISSPROT for Phylogena requires ~1GB disk space; for TrEMBL, you'll need something around 4 GBs. Then

1. decide where you want to put the sequence database and set up a directory for it. We will call this folder the database root directory in the following
2. create three further directories in your database root directory with the following names: `database`, `formatdb` and `biojava`.
3. download the fasta formatted sequence and the swissprot formatted annotation files from EBI and unpack them (using WinZip or gzip) into the directory called `database`. For swissprot, you find the files under

ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release/uniprot_sprot.dat.gz
ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release/uniprot_sprot.fasta.gz

4. run `formatdb` from the blast package to prepare a blastable database in the directory `formatdb`. To do this on Windows, start a command interpreter (go to the Start menu, click Run, type `cmd` and press Enter) and change directory (`cd`) to the database root directory you have set up point 1. Then

```
cd formatdb
```

and type

```
your_blast_path/formatdb -i ../database/uniprot_sprot.fasta -t  
uniprot_sprot -o T -n uniprot_sprot
```

(Substitute *your_blast_path* with the location of the directory containing the blast executables on your computer, like `C:/phylogena/bin/blast`).

Upon completion (after some minutes), you should see five files called `uniprot_sprot` with different extensions and one called `formatdb.log` in your `formatdb` folder.

b) Installing a fasta sequence database

Similarly to installing an EMBL / SWISSPROT format database, set up a database root folder where you will store the data. Also create the directories `formatdb` and `biojava` in your database root directory. Then format your fasta file for BLAST within the `formatdb` directory (see above by installing a SWISSPROT type DB). After this, edit your config file specifying the name of the database (this should be the same as you used with `formatdb`), its type ("fasta") and the location of its root directory (see below, "Edit the config file").

➤ Edit the config file

The config file (by default, it is called `phylogena.config`) specifies the locations and some other features of executables and databases using key – value pairs, one per line. A key is separated from its value by an = sign; any lines starting with a # are ignored by the software.

You have to specify the following attributes before being able to start Phylogena (substituting the example paths by the ones that apply to your computer):

c) the location of the Phylogena files (listed under 3) in your file system:

```
working.dir = C:/phylogena/
```

d) the locations where you installed the software required by Phylogena:

```
### BLAST ###
blast.executable = C:/phylogena/bin/blast/blastall.exe

### CLUSTALW ###
alignment.useSoftware = clustalw
clustalw.executable = C:/phylogena/bin/clustalw/clustalw.exe
# etc.

### PHYLIP ###
phylip.directory = C:/phylogena/bin/phylip/exe/
# etc.
```

In the cases of most executables, you need to specify the whole path to the executables themselves, like shown above, whereas in the case of PHYLIP, you need to specify the path to the directory in which the executables are located, like in the above example.

With some (LINUX) versions of PHYLIP, the name of the treefile written by `neighbor` appears to be `treefile` instead of `outtree`. If this is the case with your version, you'll also need to exchange

```
phylip.outtree = outtree
```

to

```
phylip.outtree = treefile
```

in the config file.

e) the location of your BLAST database (referred to as database root directory above)

```
### sequence database ###

database1.type = swissprot
database1.name = uniprot_sprot
database1.root = C:/phylogena/data/swissprot/
```

If you want to install more than one database, you will need to specify the same set of parameters for the other databases as well, only changing the number in the parameter names. E.g., for a second database (suppose you also install TREMBL):

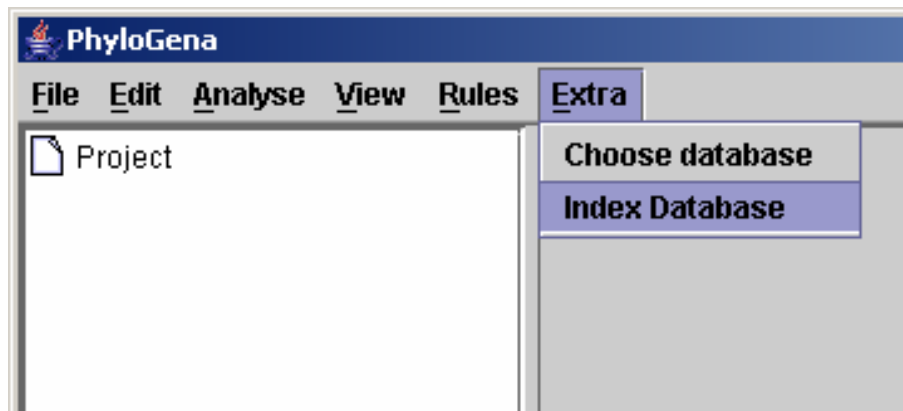
```
database2.type = swissprot
database2.name = uniprot_trembl
database2.root = C:/phylogena/data/trembl/
```

You will also have to specify how many different databases you have installed and which one of them to use as a default after startup (you can also select the database interactively in the program after you've started it, but a default has to be specified). This should look like:

```
database.number = 2
database.use = uniprot_sprot
```

➤ Run biojava indexing for each database

You need to index your database flatfile to enable Phylogena to find sequences in your local databases before you can start using the software. This simply includes choosing the database to be indexed (Extras / Choose database) and clicking on Extras / Index Database.



If nothing happens for a couple of minutes after clicking on Index Database, it is a good sign, the indexing is probably running. Wait until finished and

10. *Troubleshooting*

At the very least, you can contact me (Bank dot Beszteri at awi dot de) for the moment – I'll try my best to help you.