# PHYRN 1.5 Manual

# Draft 5/9/2011

Gaurav Bhardwaj[1], Randen L. Patterson[2], Damian B. van Rossum[3]

[1] Department of Pharmacology
University of California, Davis
Davis, CA 95616
gauravcd34@gmail.com

[2] Department of Physiology and Membrane Biology
University of California, Davis
Davis, CA 95616
rlpatterson@ucdavis.edu

[3] Department of Biology
Pennsylvania State University
State College, PA 16802
dbv10@psu.edu

## 1. Overview

PHYRN is suite of scripts that are designed to construct phylogenetic trees for highly divergent protein datasets. PHYRN uses the power of Position Specific Scoring Matrices (PSSM) and a unique composite scoring approach, to collect the meaningful signal at high divergence rates of evolution. PHYRN is designed to handle large rapidly evolving and/or highly divergent data sets. At present no binaries are are available for installation. Current version is collection of scripts that are custom-built to work at our LINUX/UNIX environment. But these scripts can easily be modified for any UNIX/LINUX based system and thus can be used on any platform.

We are currently working on making PHYRN available as a package and it should be available soon.

## 2. Download

All the scripts required for running PHYRN are available at:

**http://ccp.psu.edu/wordpress/**

## 3. Installation

PHYRN is a combined effort of multiple different biologists and software professionals. Owing to expertise of people in different scripting and coding languages, PHYRN has been written in multiple different languages. Thus, although the scripts do not need to be installed, all these languages (and some modules) need to be installed before using these scripts.

Here is the list of languages and modules that need to be installed before using these PHYRN.

1. Java (http://www.java.com/en/)
2. PYTHON (http://www.python.org/)
3. BIOPYTHON (http://biopython.org/wiki/Main_Page)
4. NUMPY (http://numpy.scipy.org/)
5. PERL (http://www.perl.org/)
6. rpsBLAST (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/rpsblast.html)
7. PSI-BLAST (http://www.ncbi.nlm.nih.gov/blast/producttable.shtml#blastp)

Depending on where you wish to implement the scripts for PHYRN (server/PC), some of these modules and languages may already be installed. So you should check installation of these languages and modules before moving further.

## 4. PHYRN METHODS

PHYRN methodology can be divided into 5 main steps, namely

1) Generation of Position Specific Scoring Matrices (PSSM)
2) Collection of alignments between query sequences and PSSM library (rpsBLAST)
3) Parsing Alignment information into Individual Summary Files
4) `Generating Composite Score Matrix`
5) `Generating Euclidean Distance Matrix`

We are going to discuss each of these steps in more detail in following sections.

### 4.1 Generation of Position Specific Scoring Matrices (PSSM)

We use a custom built PERL script (genPSSM.pl) to generate a PSSM library that can be used in rpsBLAST steps. This script utilizes PSI-BLAST to a PSSM library from a group of sequences (in fasta format)

Usually, for highly divergent protein sequences we define the boundaries of homologous regions using Conserved Domain database, Interproscan and/or literature. Once the boundaries are defined, homologous regions are extracted and stored in a separate file (pssm_fragments_file) in fasta format. 'genPSSM.pl' converts this pssm_fragments_file into a library of pssms.

Usage:

```
./genPSSM.pl [Query File] [DB name] [Db description]
[working directory path] [PSSM start number]
```

**Query file**: Fasta file with all the chopped fragments (homologous fragments)
**DB name**: Name you wish to give to your PSSM library
**DB description**: Any identifier that you want to associate with the new PSSM library (e.g expanded_binding library
**PSSM start number**: Numbering of PSSM is going to start from here (e.g 1 if you want to start your first pssm to be named pssm0001)

'genPSSM.pl' generates a name_masters.fa file which is the sequence file for PSSMs generated. But these are not the actual PSSMs. Actual PSSMs are generated into two different formats: 'PSSM' directory contains individual PSSM files, while 'CDD' directory contains the compiled PSSM library. We use the compiled PSSM library for running rpsBLAST.

### 4.2 Collection of alignments between query sequences and PSSM library

We use rpsBLAST to collect all possible alignments between Full-Length protein sequences (belonging to same protein family) and PSSM library (generated from the homologous regions of the protein). All the alignments are collected and tabular data corresponding to the identity and coverage is stored in a single output file (rpsBLAST output file)

USAGE:

**blast-2.2.18/bin/rpsblast -i [Query File] -d [Database path] -e [e-value threshold] -m 8 -o [output file name]**

e-value threshold can be selected across range of values starting from very stringent (0.01) to lenient (10000000000). For evolutionary measurements, tests with simulated datasets show that lenient thresholds work better as we collect all possible data points, which later get screened out based on identity and coverage. But lenient threshold only work better if PSSM library is well defined and is of good quality. If the PSSM library has noise from non-homologous regions, lenient e-value threshold can add noise to the measurements, translating into a phylogenetic tree of poor quality and/or less robustness.

### 4.3 Parsing Alignment information into Individual Summary Files

rpsBLAST generates a single file with all the alignment information between all possible query-PSSM pairs. For calculating the composite score matrix, we generate individual summary files for each query. These text files contain tabular information of alignments obtained from each full-length query sequence and members of PSSM library. We use '**genSUMMARY.jar'** script for parsing the rpsblast output file to individual summary files.

USAGE:

**java –jar genSUMMARY.jar [Query file] [PSSM masters file] [Output file] [Output directory for generating inborns]**

NOTE: Output directory needs to be generated manually before running inborn.jar (in Unix environment it can be done using mkdir)

### 4.4 Generating Composite Score Matrix

'genCSMAT.py' calculates composite score (identity X coverage) between each query-PSSM pair, and converts this information into a N X M composite score matrix, where N is the number of queries and M is the number of PSSMs.

USAGE:

**python genCSMAT.py –p [PSSM masters file] –q [Query file] –suffix .txt –f [inborn folder]**

NOTE: Suffix by default is .txt, but needs to be changed if summary files are generated with a different extension.

genCSMAT.py outputs the composite score matrix in a file with .tab extension.

### 4.5 Generating Euclidean Distance Matrix

'genEUCLID.py' converts a composite score matrix into an Euclidean distance matrix. The output distance matrix format is MEGA, so the output Euclidean distance can easily be opened with MEGA, to calculate phylogenetic tree using Neighbor-Joining (NJ) and/or Minimum Evolution (ME).

USAGE:

```
python genEUCLID.py [N X M matrix {tab} file]
```