

# Accelerating Protein Phylogenetic Analysis by PHYLIP on NUS Grid

Hu Yongli

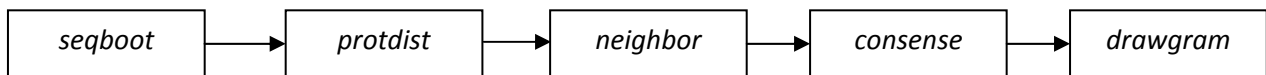
(Dept. of Biochemistry, Yong Loo Lin School of Medicine)

## Introduction

PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). Developed in the 1980s, PHYLIP is one of the most widely-distributed phylogeny packages, and it has been used to build the largest number of published trees. Currently, PHYLIP has over 15,000 registered users worldwide. The PHYLIP package is available free over the Internet <sup>(2)</sup>, and has been written to work on many different kinds of computer systems such as Windows, MacOS and Linux systems. Methods included in the package include parsimony, distance matrix, and likelihood methods. Particularly important is the bootstrapping and consensus trees programs which are vital tools for scientists to build phylogentic trees for inference of evolutionary events. PHYLIP is also able to handle different data types which include Deoxyribonucleic Acid (DNA) and protein sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters <sup>(1)</sup> to suit different users with their specific needs for phylogentic analyses.

## Grid Enabling of PHYLIP

This report will explain how PHYLIP can be customised and be placed on the Grid to aid in the acceleration of conventional protein phylogenetic tree construction. The typical steps undertaken in building a phylogenetic tree with protein sequence information is shown in Figure 1. First, a protein sequence alignment file (presented in .phy format) is placed into the program *seqboot* in PHYLIP, which reads in the input data set and produces multiple data sets (typically 100) by bootstrap resampling of the input data. Next, the output from *seqboot* is subsequently used by the program *protdist* which computes a distance measure for protein sequences required for downstream tree building. In this step, each of the 100 *seqboot* output datasets will be discretely used for the calculation of protein distances and this unique property allows the serialisation of the processing of data by *protdist* as the *protdist* outputs from one *seqboot* input dataset is not required for the processing of the next *seqboot* dataset and allows for data-processing on multiple machines on the grid. Subsequently, output generated from *protdist* will be used by the program *neighbor*, where the clustering of sequences within each individual dataset can be carried out before merging the *neighbor* datasets for the final program *consense* where a consensus tree can be built with the 100 datasets.



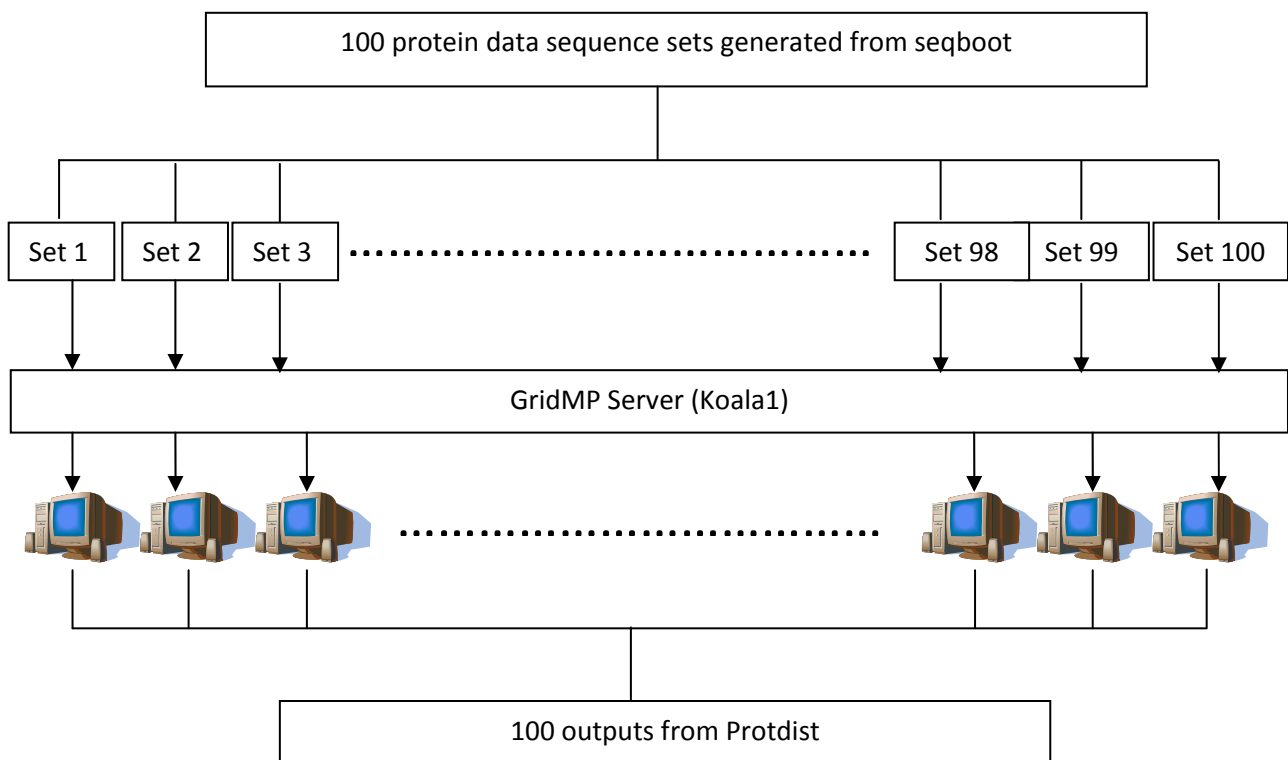
**Figure 1:** Typical pipeline for building a protein phylogenetic tree with PHYLIP. The output of one program is used as the input for the next program. For example, the output for *seqboot* is used as the input for the program *protdist* and the output from the program *protdist* is used by *neighbor*.

To enable PHYLIP on the grid, it is important to identify the step(s) in the protein phylogenetic tree construction pipeline that allows the processing of the data to be broken up into small portions. The manipulation of each portion of the data must be done independently of the other datasets. Taking into

account these criteria, the most ideal candidate for the parallelisation process is data processing by program *protdist*. Other than leveraging on the technical advantage of parallelising *protdist*, the selection of *protdist* is also based on preliminary analyses which show that *protdist* takes the longest running time of 9 hours and 6 minutes on a dataset of 178 protein sequences on a Sunfire 6800 server with 16 CPUs at 900MHz and 16GB RAM, while other programs like *seqboot*, *neighbor* and *consense* each takes less than two minutes to complete.

Taking advantage of the ability to individually run the *protdist* program with each *seqboot* output, the grid-enabled *protdist* (with the application called meta-PHYLIP) will take the 100 split *seqboot* outputs and distribute them to 100 machines for processing. The steps undertaken to enable protein sequence data analysis is as shown in Figure 2.

1. a *seqboot* output file, containing 100 datasets, will be generated from the protein sequence alignment.
2. the *seqboot* output file will be divided into 100 files, each file containing one of the 100 outputs generated by *seqboot*.
3. the files will then be uploaded to GridMP server (koala1), running the Univa UD Grid MP platform middleware, which will dispatch the input data file, together with the parameter files and the *protdist* program, to the available grid client machines for processing.
4. upon completion, the output will be sent back to koala1 for the user to download.



**Figure 2:** A simplified schematic representation of the flow of data for grid-enabling of data processing by software *seqboot*.

## Results

Preliminary results on running meta-PHYLIP on NUS TCG grid have achieved a speed up of 24 to 58 times as compared to running *protdist* on a standalone machine. The latest 10 jobs processed on NUS TCG Grid show a very constant speedup rate of more than 50, see Table 1.

**Table 1:** Latest 10 PHYLIP Jobs Processed NUS TCG Grid

No	JOBID	Start Time	End Time	Running Time (Minutes)	CPU Time	CPU Minutes	Speedup Rate
*1	10220	9/23/2009 6:33	9/25/2009 3:23	2,690	105d13h35m	152,015	56.5
2	10219	9/21/2009 16:45	9/22/2009 22:00	1,785	68d8h9m	98,409	55.1
3	10218	9/20/2009 21:45	9/21/2009 13:36	951	38d17h2m	55,742	58.6
4	10217	9/20/2009 11:31	9/20/2009 19:31	480	17d2h35m	24,635	51.3
5	10216	9/20/2009 11:24	9/20/2009 13:15	111	4d7h58m	6,238	56.2
6	10215	9/18/2009 10:19	9/18/2009 10:50	31	1d1h51m	1,551	50.0
*7	10140	9/3/2009 9:57	9/11/2009 2:54	11,097	386d12h18m	556,578	50.2
8	10078	8/27/2009 7:37	8/28/2009 22:10	2,313	84d11h46m	121,666	52.6
9	10060	8/25/2009 13:38	8/26/2009 13:16	1,418	53d20h6m	77,526	54.7
10	10041	8/23/2009 10:21	8/23/2009 12:51	150	5d10h4m	7,804	52.0

\*The rows colored in blue (Job No 1 and 7) show the jobs that will require more than three months to complete if these processes were run on standalone machines.

It should be highlighted that for jobs No. 1 and 7 which consumed about 3½ months and 12.7 months CPU time on Grid, it will be not practical to perform the sequences computation on a workstation as the completion time will be too long and incur unnecessarily long run time and render the analysis of large dataset impossible. Enabling *protdist* in meta-PHYLIP on the grid will greatly enhance the process of protein phylogentic tree construction and contribute positively to biological research.

## Acknowledgements

I would like to thank A/P Tan Tin Wee for conceiving the idea of parallelising PHYLI on grid. I also wish to express my gratitude to Mr Mark De Silva and Mr Lim Kuan Siong for providing technical support, Mr Mohammad Asif Khan and Miss Heiny Tan for providing their Dengue virus and Influenza A protein sequence alignments, and also Mr Wang Junhong for his valuable comments and suggestions for the writeup.

## References

- (1) General Description of PHYLIP. <http://evolution.genetics.washington.edu/phylip/general.html> , retrieved on 14 September 2009
- (2) Porting PHYLIP phylogenetic package on the Desktop GRID platform XtremWeb-CH. (2007) *Stud Health Techno Inform.* **126**:55-64.