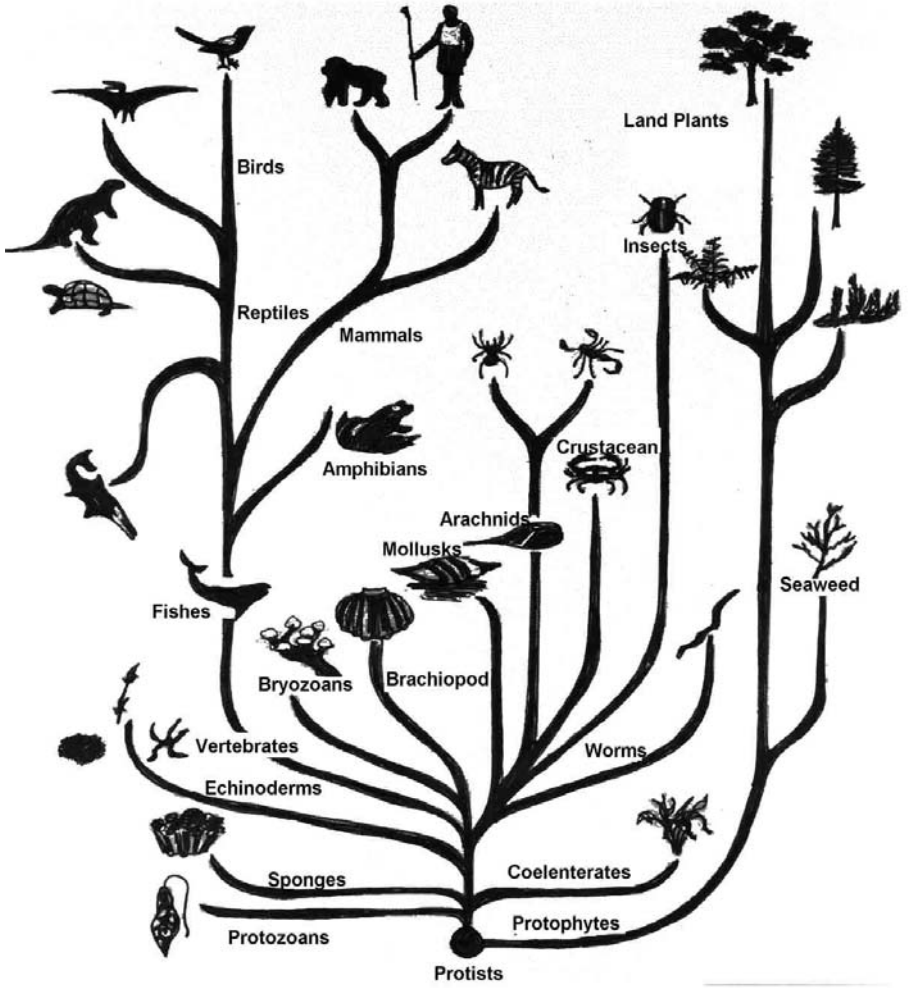
EVOLUTIONARY MOLECULAR PHYLOGENETIC ANALYSES ON THE NUS TCG GRID

Hu Yongli

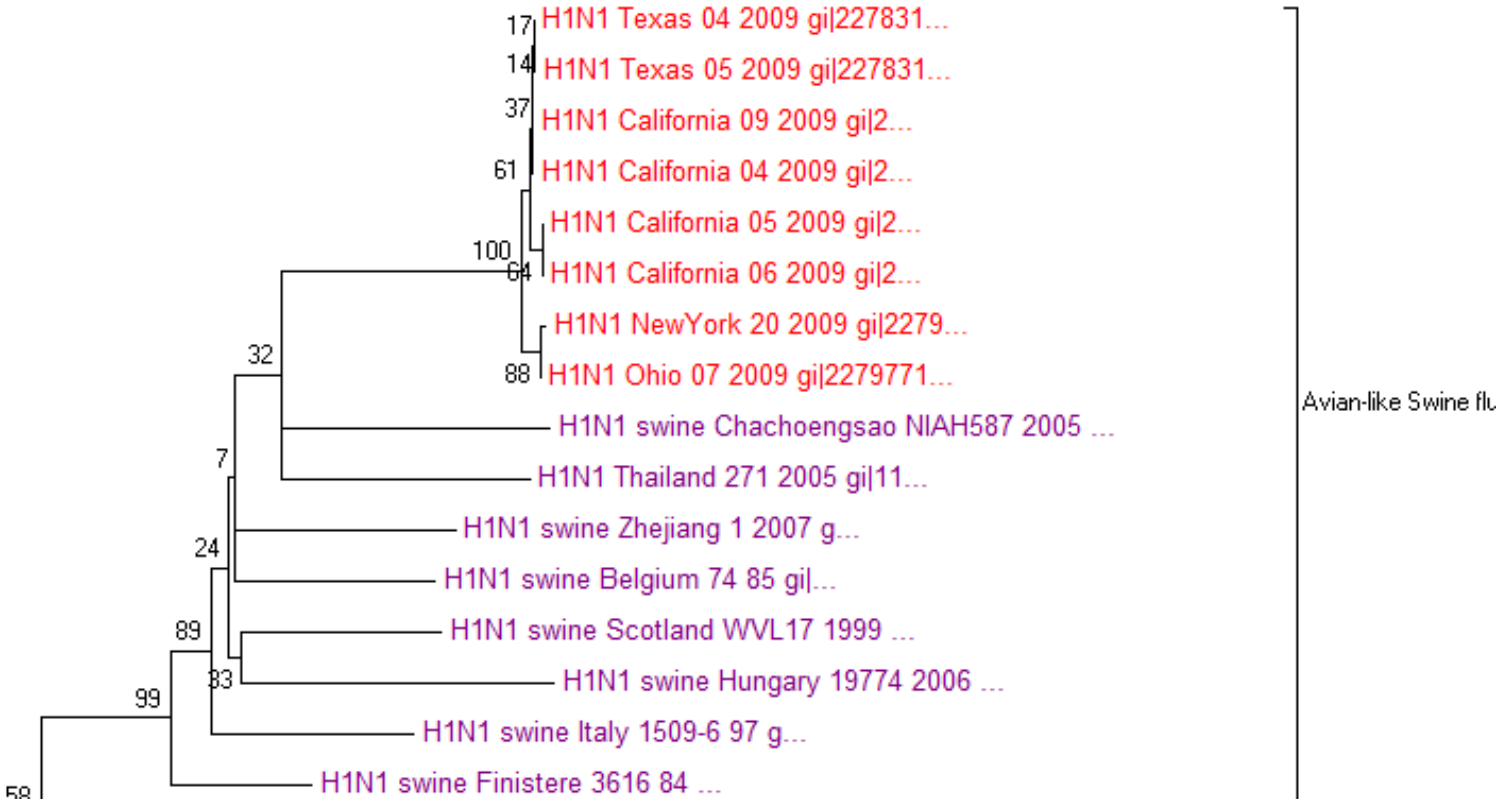
Department of Biochemistry,
Yong Loo Lin School of Medicine

WHAT IS MOLECULAR PHYLOGENY?

- The Science of estimating the evolutionary past
  - Fossil data
  - Morphological data
  - Protein sequence data
  - DNA sequence data
  - Etc…



Baldauf, S.L., 2003,*Trends Genet.* **16**(6):345 - 51
http://www.clarifyingchristianity.com/images/philotr1.gif, retrieved on 21 Nov 09



Maurer-Stroh, S. *et. al*, 2009, *Bio. Direct* **4:**18

PHYLIP 

- Developed in the 1980s
- Most commonly used package for inferring phylogenies
- Most widely - distributed phylogeny packages
- Used for building the largest number of published phylogenetic trees
- Contains a large number of methods and can handle many type of data
- Open source 

http://evolution.genetics.washington.edu/phylip/general.html, retrieved on 21 Nov 09
Abdennadher, N. and Boesch, R. , 2007, *Stud Health Technol Inform*. **126**:55 - 64

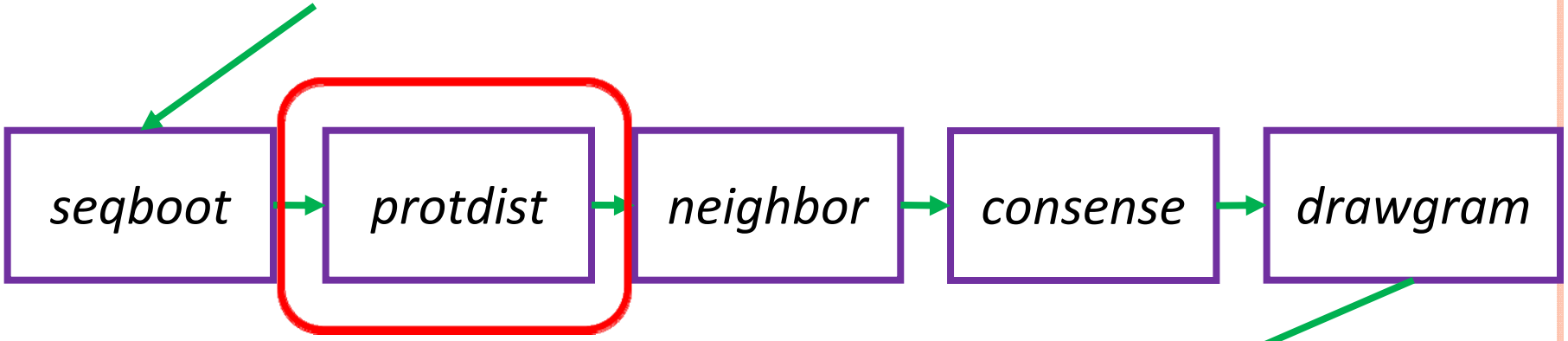
BUILDING A PROTEIN PHYLOGENETIC TREE

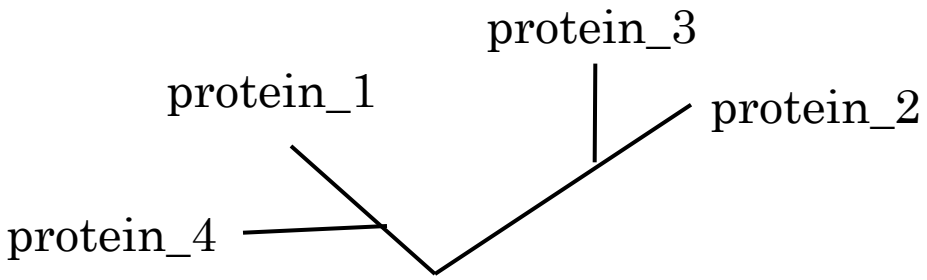
>protein_1
GJYWLKADWWGGMD…

>protein_2
KKLLDWGGJWGGMD…

>protein_3
KKLLDWGKJWGGME…

>protein_4
GJYWLAADWWGGMS…





WHY PROTDIST???

- Most time consuming step
  - Building a tree with 178 protein sequences *
  - protdist ~9 hours and 6 minutes
  - seqboot, neighbor and consense ~ 2 minutes each

- Ability to be parallelized to be placed on the grid
  - each of the 100 *seqboot output datasets can be discretely used for the* calculation of protein distances in *protdist*

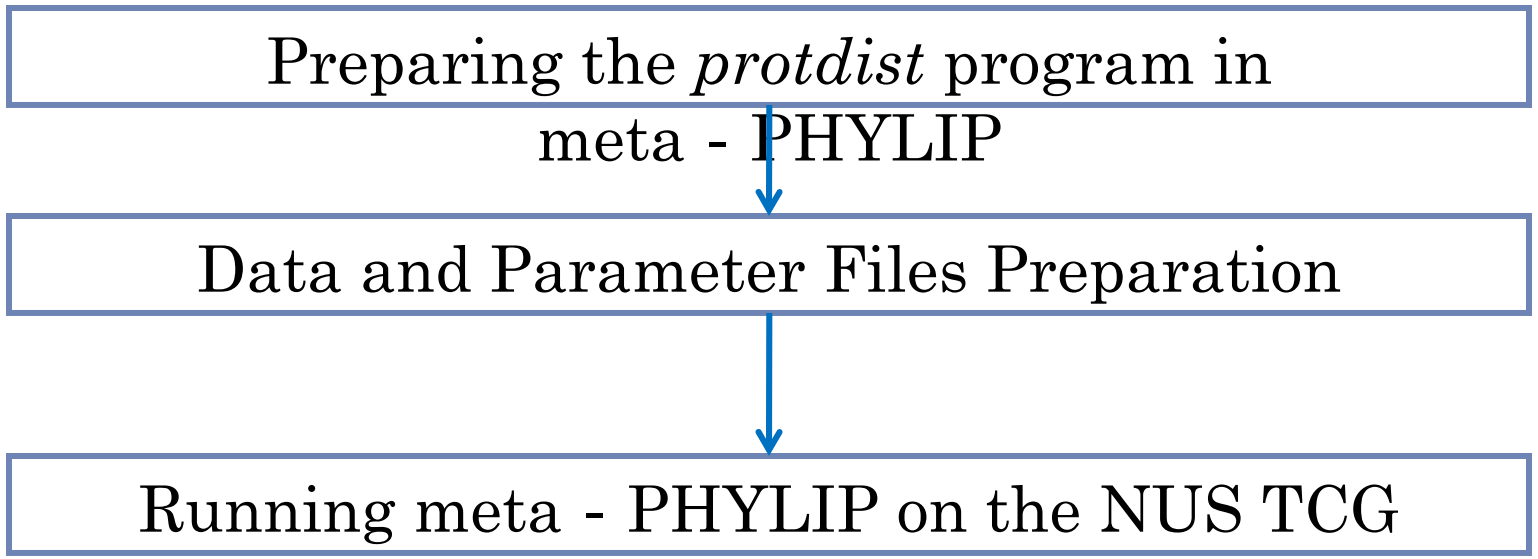
*Sunfire 6800 server, with 16 CPUs at 900MHz and 16GB RAM



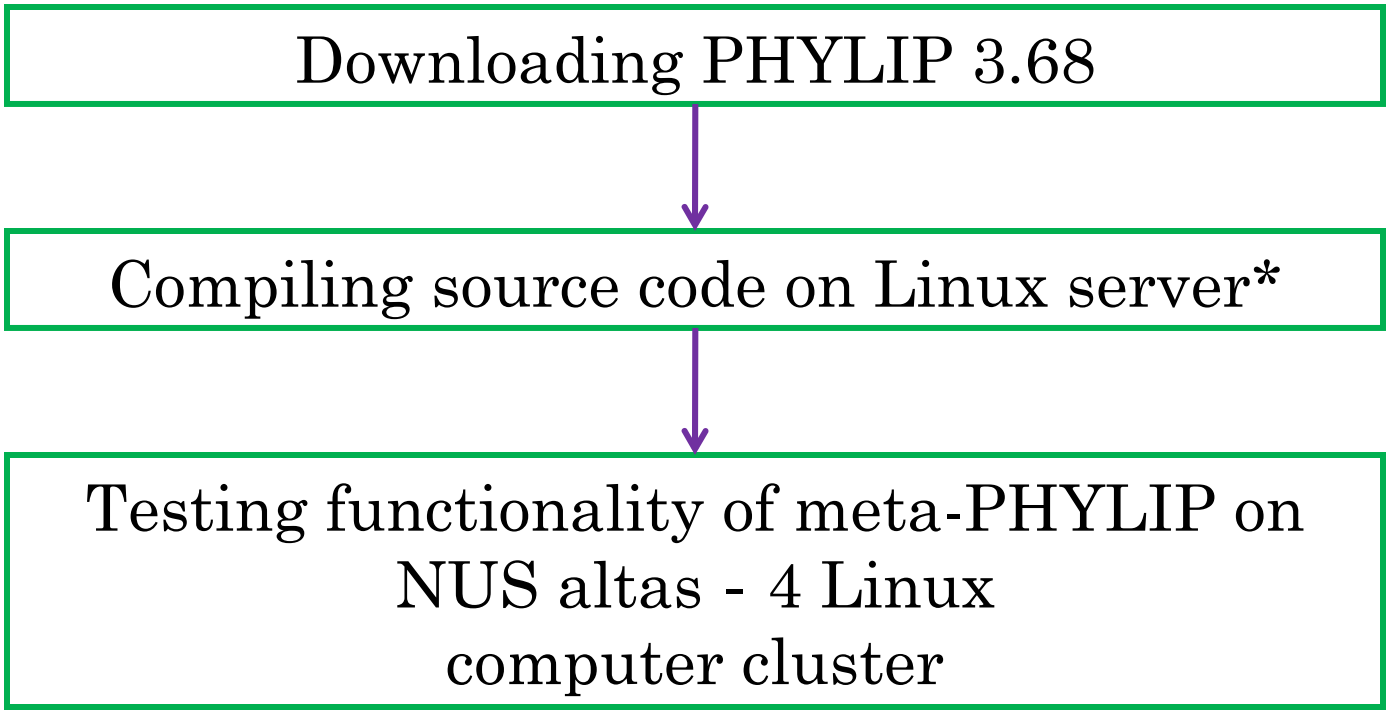
ENABLING PHYLIP ON NUS TCG



STEPS TAKEN TO PLACE META-PHYLIP ON NUS TCG

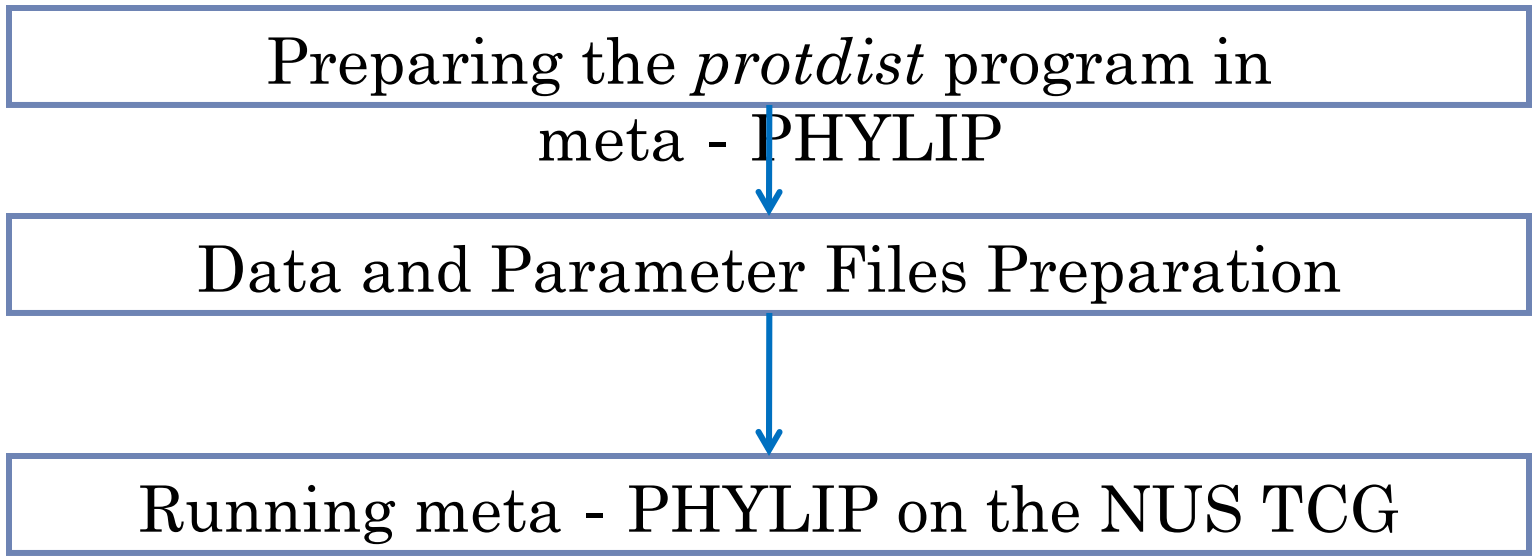


PREPARING THE *PROTDIST* PROGRAM IN META - PHYLIP



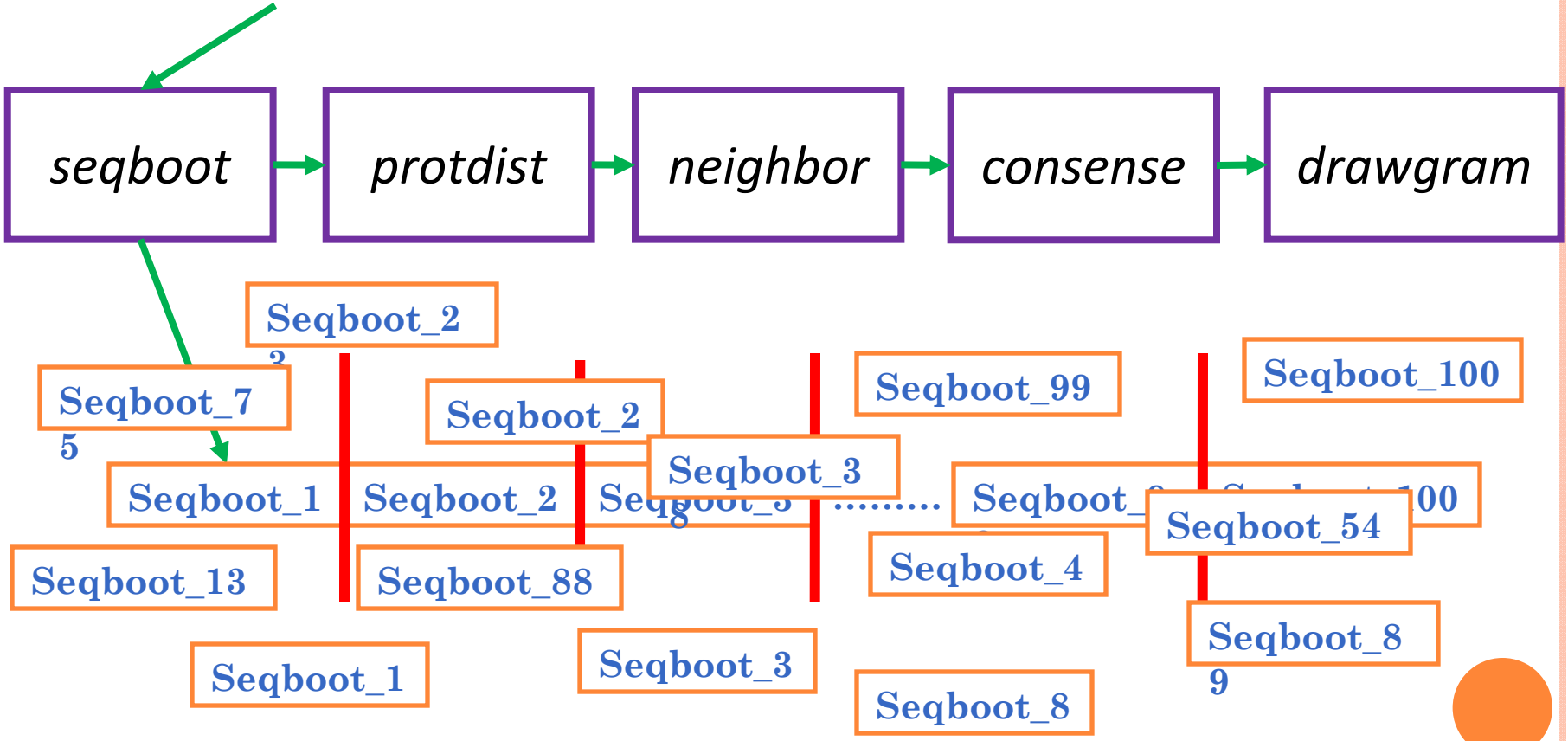
* Intel Pentium 4 CPU 3.00GHz, 4 GB of RAM running on Slackware 10.0

STEPS TAKEN TO PLACE META-PHYLIP ON NUS TCG GRID



DATA AND PARAMETER FILE PREPARATION (DATA FILES = INPUT1.DAT)

>protein_1
GJYWLKADWWGGMD…
>protein_2
KKLLDWGGJWGGMD…
>protein_3
KKLLDWGKJWGGME…
>protein_4
GJYWLAADWWGGMS…



DATA AND PARAMETER FILE PREPARATION (PARAMETER FILES = INPUT2.DAT)

Parameter File

input1.dat
F
output1.dat
Y

```
protdist.exe: can't find input file "infile"
Please enter a new file name> input1.dat

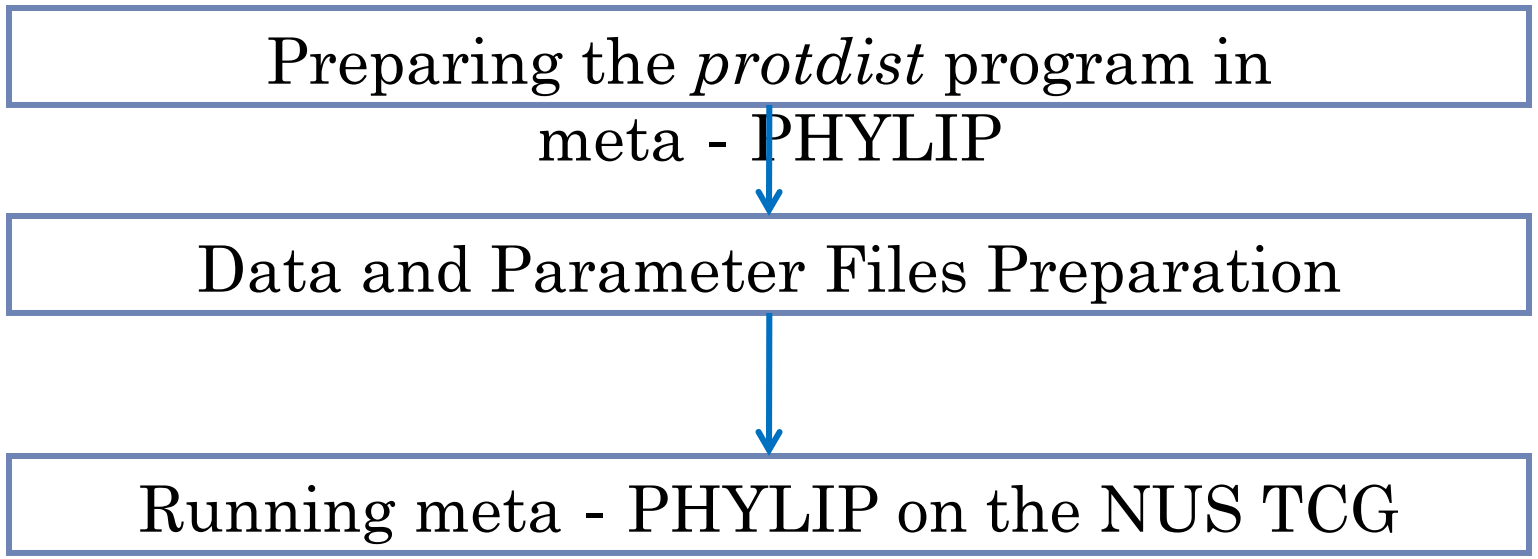
protdist.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
F
Please enter a new file name> output1.dat

Protein distance algorithm, version 3.69

Settings for this run:
  P  Use JTT, PMB, PAM, Kimura, categories model?  Jones-Taylor-Thornton matrix
  G  Gamma distribution of rates among positions?  No
  C            One category of substitution rates?  Yes
  W                     Use weights for positions?  No
  M                   Analyze multiple data sets?  No
  I                  Input sequences interleaved?  Yes
  0                Terminal type (IBM PC, ANSI)?  IBM PC
  1            Print out the data at start of run  No
  2          Print indications of progress of run  Yes

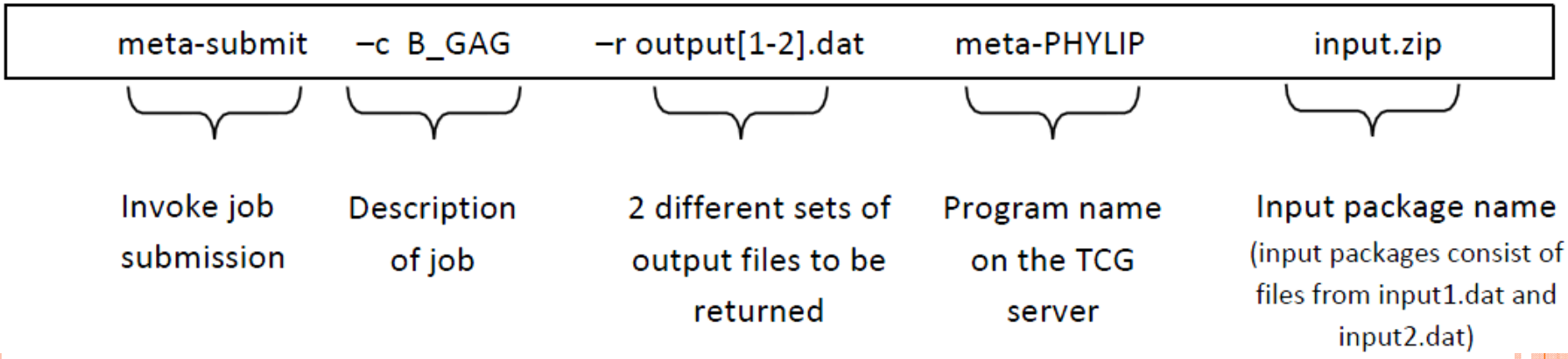
Are these settings correct? (type Y or the letter for one to change)
Y
```

STEPS TAKEN TO PLACE META-PHYLIP ON NUS TCG



RUNNING META - PHYLIP ON THE NUS TCG

- Download parametrics study program
- Prepare zipped input file: "input.zip" (data+parameter files)



DATA PROCESSING ON GRID

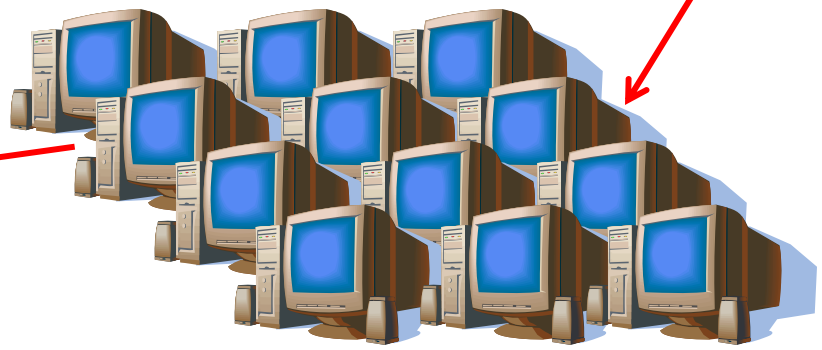
Input.zip
(100 seqboot output files + 100 parameter files)

Koala1
(GridMP Server)



Seqboot_1 — Param_1 — Meta-PHYLIP
Seqboot_2 — Param_2 — Meta-PHYLIP
Seqboot_3 — Param_3 — Meta-PHYLIP
Seqboot_99 — Param_99 — Meta-PHYLIP
Seqboot_100 — Param_100 — Meta-PHYLIP

Output1.dat.000100
Output2.dat.000100
Output1.dat.000001
Output2.dat.000001
Output1.dat.000002
Output2.dat.000002
Output1.dat.000099
Output2.dat.000099



LOG FILES

Parameter File

input1.dat
F
output1.dat
Y

```
protdist.exe: can't find input file "infile"
Please enter a new file name> input1.dat

protdist.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
F
Please enter a new file name> output1.dat_

Protein distance algorithm, version 3.69

Settings for this run:
  P  Use JTT, PMB, PAM, Kimura, categories model?  Jones-Taylor-Thornton matrix
  G  Gamma distribution of rates among positions?  No
  C            One category of substitution rates?  Yes
  W                     Use weights for positions?  No
  M                   Analyze multiple data sets?  No
  I                  Input sequences interleaved?  Yes
  0                Terminal type (IBM PC, ANSI)?  IBM PC
  1            Print out the data at start of run  No
  2          Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
Y_
```

EVALUATING THE SPEEDUP OF META-PHYLIP



EVALUATION OF SPEEDUP

$$\textbf{Speedup} = RT_{100} / Tp$$

RT_{100} : time (in seconds) from the job creation to return of the last output to the grid server

Tp : total CPU time required to run the program in serial.

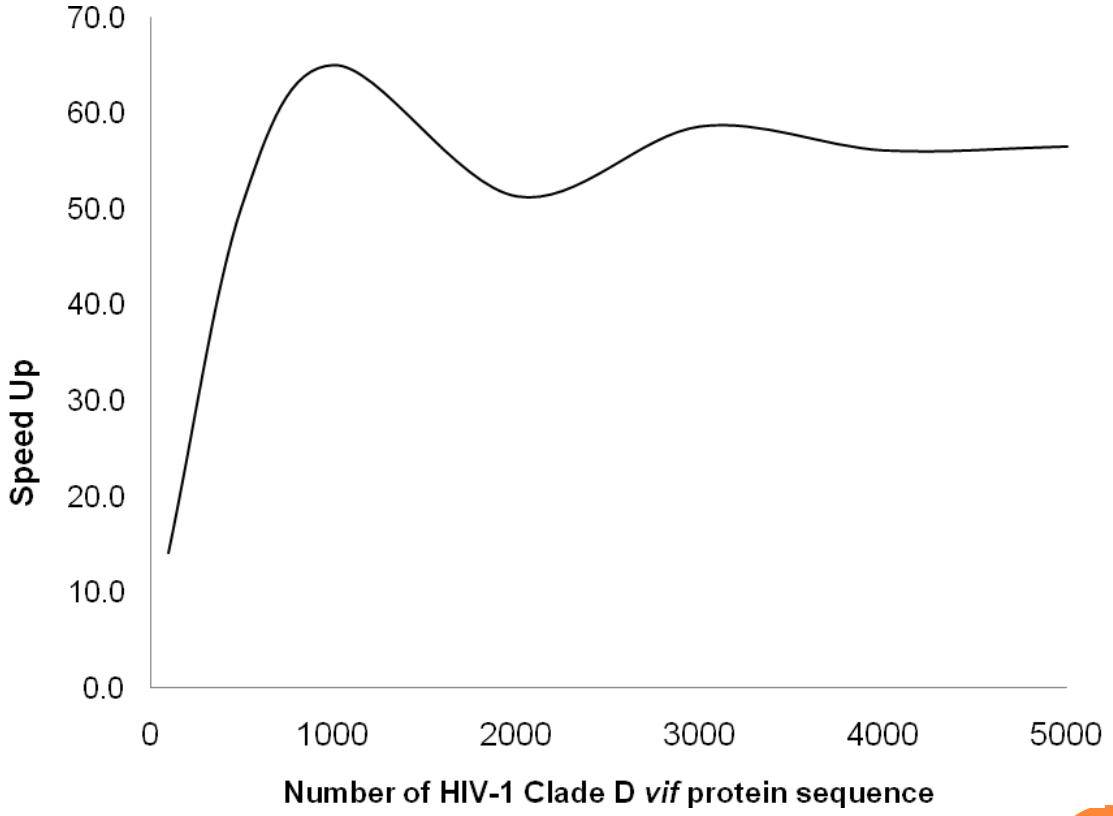
Speedup is explored with

- Same protein length different number of protein sequences
- Real-life biological datasets



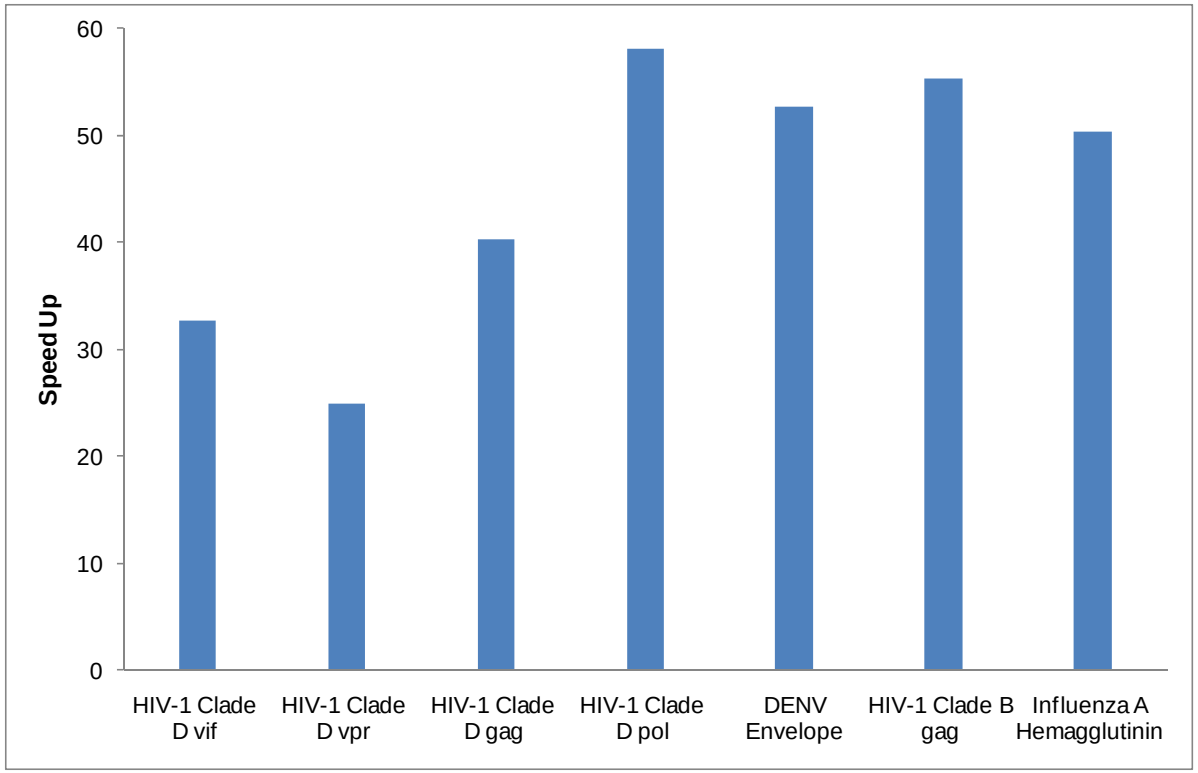
SPEEDUP ACHIEVED WITH DATASET OF DIFFERENT NUMBER OF SEQUENCES

- speedup achieved ranges from 14.1 to 65.0 times
- speedup for small datasets is lower than larger datasets



SPEEDUP ACHIEVED WITH REAL BIOLOGICAL DATA

- speedup achieved ranges from 25.0 to 58.1 times
- speedup for small datasets is lower than larger datasets



DISCUSSION AND CONCLUSION

- Advancement in sequencing technology brings about sequence data explosion
- Phylogenetic analyses can no longer be carried out within an acceptable time frame
- Placing PHYLIP on the grid will greatly enhance the rate of molecular phylogenetic analyses
  - Acceleration depends on availability of idle computer cycles on grid clients
- Importance in the study of disease outbreaks and emerging pandemics, especially in disease treatment and pandemic containment

Sanderson, M.J. and Driskell, A.C. ,2003, Trends Plant Sci. 8(8):374 - 379
Maurer-Stroh, S. et. al, 2009, Bio. Direct 4:18

ACKNOWLEDGEMENTS

- A/Prof Tan Tin Wee
- Mark De Silva
- Lim Kuan Siong
- Wang Jun Hong
- Mohammad Asif Khan
- Heiny Tan
- All members of BIC

THANK YOU