PhyASBT: Phylogenetic Sequence Alignment, Model Selection, Bootstrapping and Tree Viewer

Tzu-Cheng Chuang, Shen-Yao Su, Chung-Yen Lin and Jan-Jan Wu

Institute of Information Science, Academia Sinica, Taipei, Taiwan

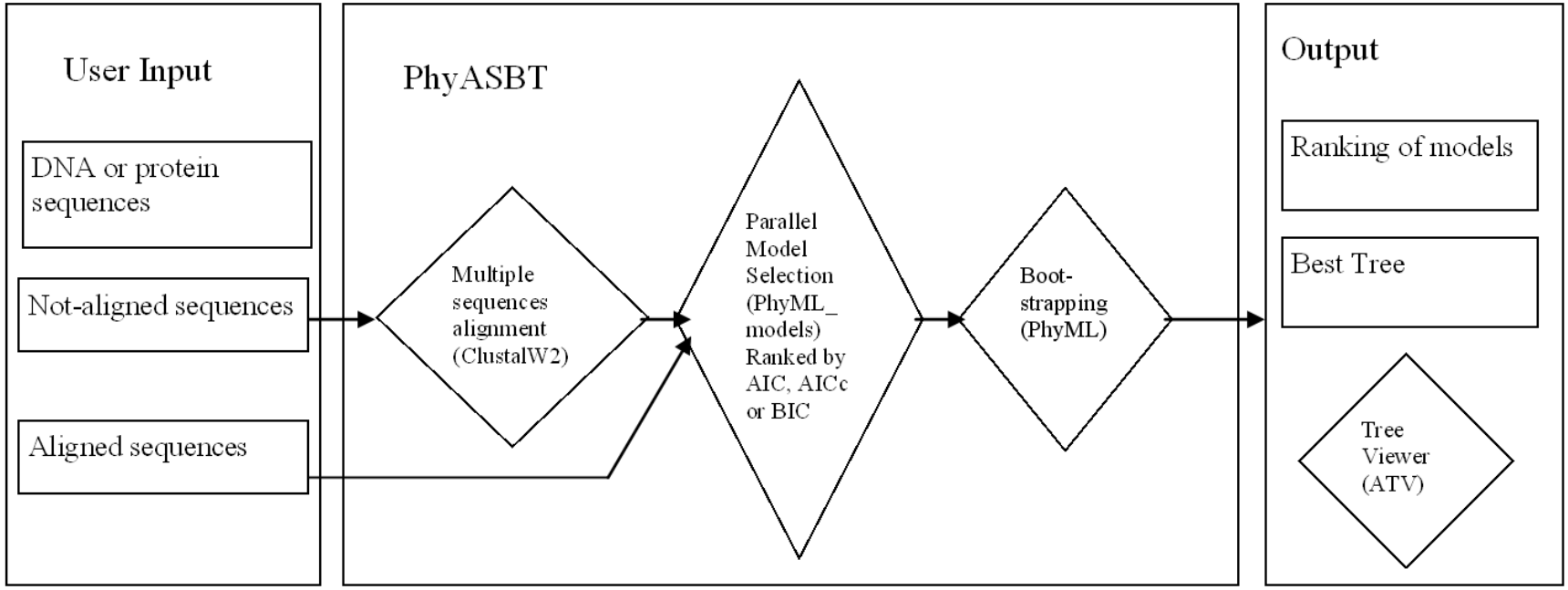
Outline

- Workflow of PhyASBT
- Model files for DNA and protein
- Desktop version
- Server version
- Options for users
- Job daemon
- Techniques
- Results from parallel program

PhyASBT: Phylogenetic Alignment Selection Bootstrap Treeviewer

- 1. Multiple Sequence Alignment:
- Clustalw2 is used for multiple sequence alignment if the sequnce is not aligned.(MPI)
- 2. Model Selection:
- PhyML is used for model selection. Users can add their own models. (OpenMP, MPI, MPI+OpenMP)
- 56 models for DNA sequences.
- 112 models for protein sequences.
- 3. Bootstrapping:
- PhyML is used for bootstrapping.(MPI)
- 4.Tree Viewer:
- ATV is used for viewing the tree.

Workflow


User Input
DNA or protein sequences
Not-aligned sequences
Aligned sequences
PhyASBT
Multiple sequences alignment (ClustalW2)
Parallel Model Selection (PhyML_ models) Ranked by AIC, AICc or BIC
Boot-strapping (PhyML)
Output
Ranking of models
Best Tree
Tree Viewer (ATV)

DNA Models file

- DNA models(dnamodel.csv): 56
- Model name, parameters, number of parameters
- JC, -m JC69 -f d -t e -v 0 -a 1, 0
- JC+I, -m JC69 -f d -t e -v e -a 1, 1
- JC+G, -m JC69 -f d -t e -v 0 -a e, 1
- JC+I+G, -m JC69 -f d -t e -v e -a e, 2
- K80, -m K80 -f d -t e -v 0 -a 1, 1
- F81, -m F81 -f d -t e -v 0 -a 1, 3

Protein models file

- Protein models(proteinmodel.csv): 112
- Model name, parameters, extra number of parameters
- JTT, -m JTT -f m -v 0 -a 1, 0
- JTT+G, -m JTT -f m -v 0 -a e, 1
- JTT+G+F, -m JTT -f e -v 0 -a e, 20
- Base number of parameters: 2*Nseq-3
- Total number of parameters=Base+Extra

Desktop version of PhyASBT

- Multiple Sequence alignment(sequential code)
- Model selection(parallel code)
- Windows and Linux versions can be run in parallel.
- Bootstrapping (sequential code)
- It’s not easy to find and install MPI compiler and MPIexec program for desktop computers.
- Desktop computer is more suitable with share-memory architecture(OpenMP)
- This whole package can be executed on Windows, Linux and MacOS.

Server version of PhyASBT

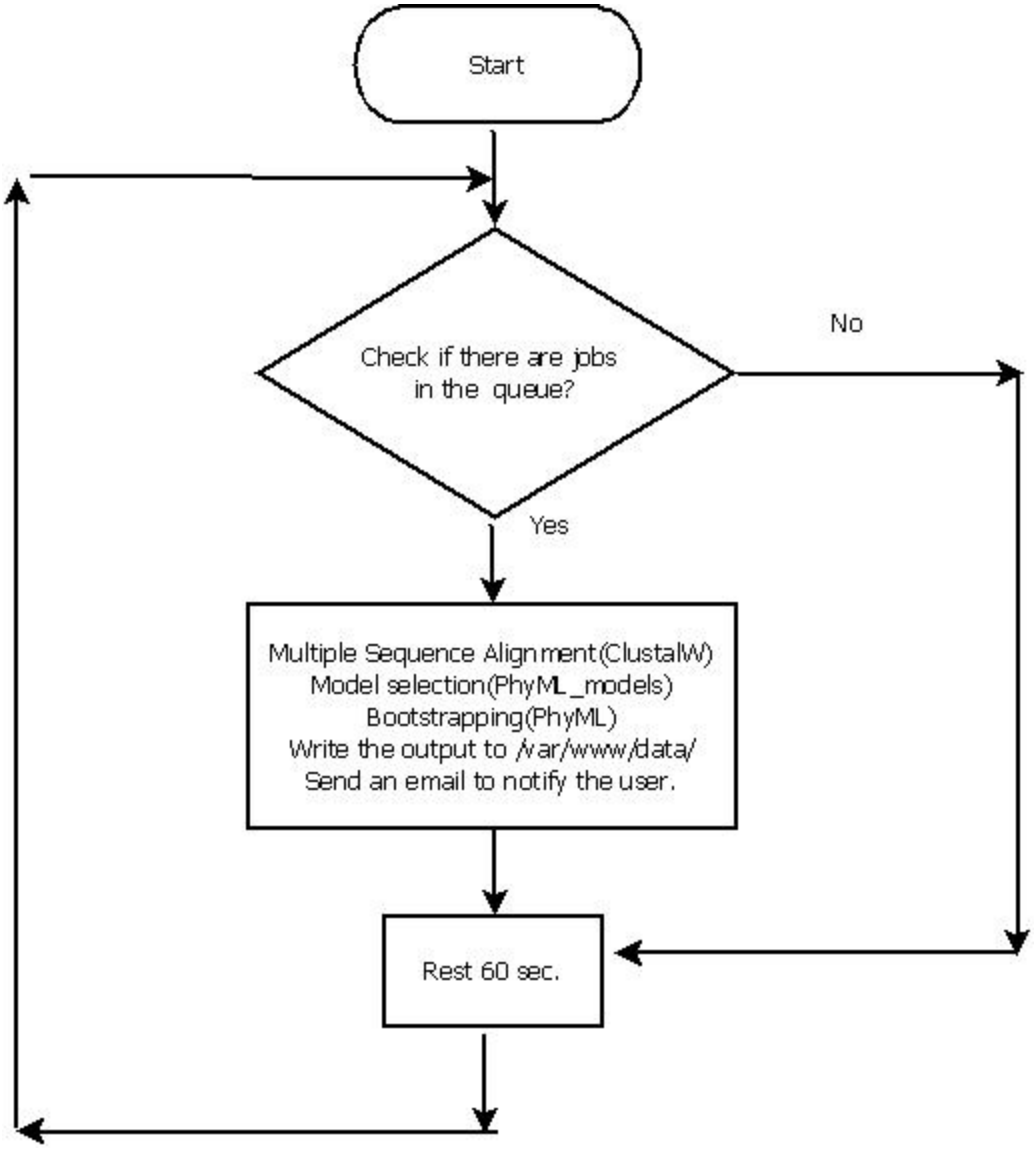
- It’s installed on Ubuntu Linux. (motif.iis.sinica.edu.tw)
- Multiple Sequence alignment(sequential code)
- Model selection(parallel code, OpenMP, MPI, MPI+OpenMP)
- Bootstrapping (sequential code)
- Treeviewer

Options for users

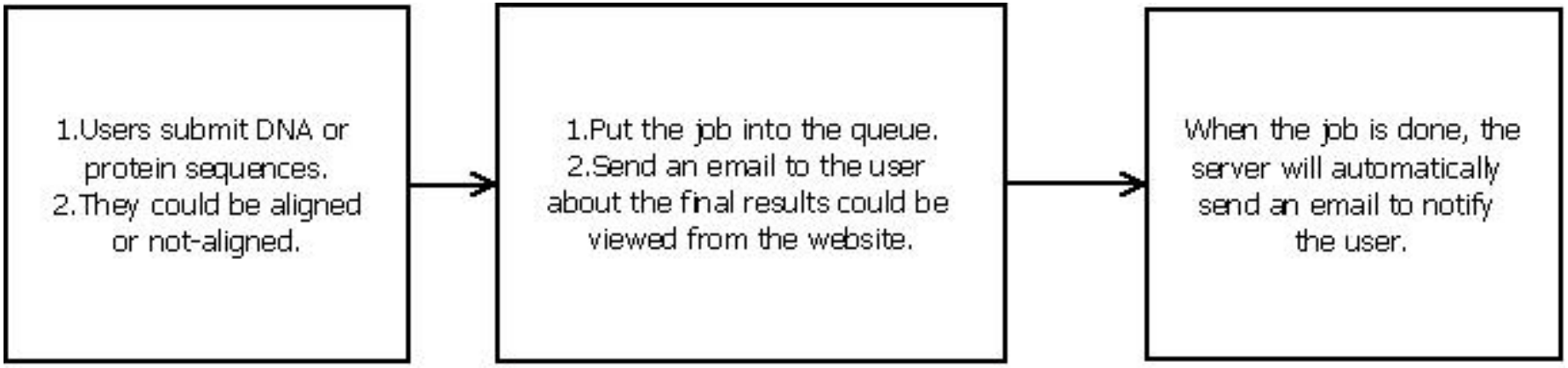
- Input sequence file(Must be given)
- Is_Aligned? (If it's aligned, please use Phylip format)
- Parameter d: aa (Protein seq.), nt (DNA seq.)
- Parameter c: number of substitution rate
- Parameter o: tlr, tl, lr, l, r, n (optimization)
- Parameter s: sorted by 1(AIC), 2(AICc), 3(BIC)
- Parameter b: number of bootstrap

Job daemon on the server

- Written in Python.
- The code should be executed when the server is started.
- It uses ‘sendmail’ to send the plaintext email.



User submits his job



Techniques

- Apache PHP server is used to receive user's data and store the information on the database.
- Python is used to control the job queue and send the email.
- OpenMP is used to reduce the computation time and balance the load.
- Computation intensive codes are written in C++.
- Java is used to design GUI for desktop version.
- When MPI is installed on the computer, clustalw, phyml bootstrap can be run in parallel.

Execution time between OpenMP and MPI (Shared memory architecture)

	# of processors	1	8	ratio
OpenMP	DNAsmall.phy -o tl	5m56s	49s	7.265
MPI	DNAsmall.phy -o tl	8m36s	1m7s	7.701
OpenMP	Pseqsmall.phy -o tl	15m45s	2m1s	7.81
MPI	Pseqsmall.phy -o tl	20m53s	2m41s	7.783
OpenMP	DNAbig.phy -o tl	234m49s	33m9s	7.083
MPI	DNAbig.phy -o tl	438m50s	56m52s	7.717
OpenMP	Pseqbig.phy -o tl	1577m14s	206m44s	7.629
MPI	Pseqbig.phy -o tl	2343m45s	294m57s	7.946
OpenMP	DNAsmall.phy -o n	54s	8s	6.75
MPI	DNAsmall.phy -o n	1m47s	16s	6.6875
OpenMP	Pseqsmall.phy -o n	3m2s	25s	7.28
MPI	Pseqsmall.phy -o n	4m52s	41s	7.122
OpenMP	DNAbig.phy -o n	53m51s	8m28s	6.36
MPI	DNAbig.phy -o n	151m57s	20m30s	7.412
OpenMP	Pseqbig.phy -o n	189m21s	27m51s	6.799
MPI	Pseqbig.phy -o n	352m36s	44m56s	7.847

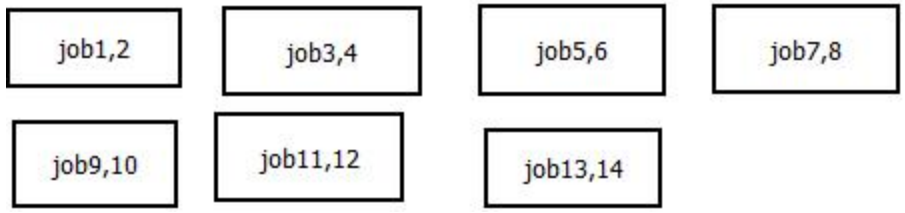
- MPI takes longer time than OpenMP.
- MPI has better speedup ratio.
- When the computation is CPU intensive, the improvement ratio is better.

MPI structure (Distributed memory architecture)

- Master CPU: distribute jobs to slave CPUs, ReceiveMessage->Send jobs to slave CPUs
- Slave CPU:Request jobs from Master CPU SendMessage to request jobs->Receive jobs from Master CPU->Compute

MPI+OpenMP structure

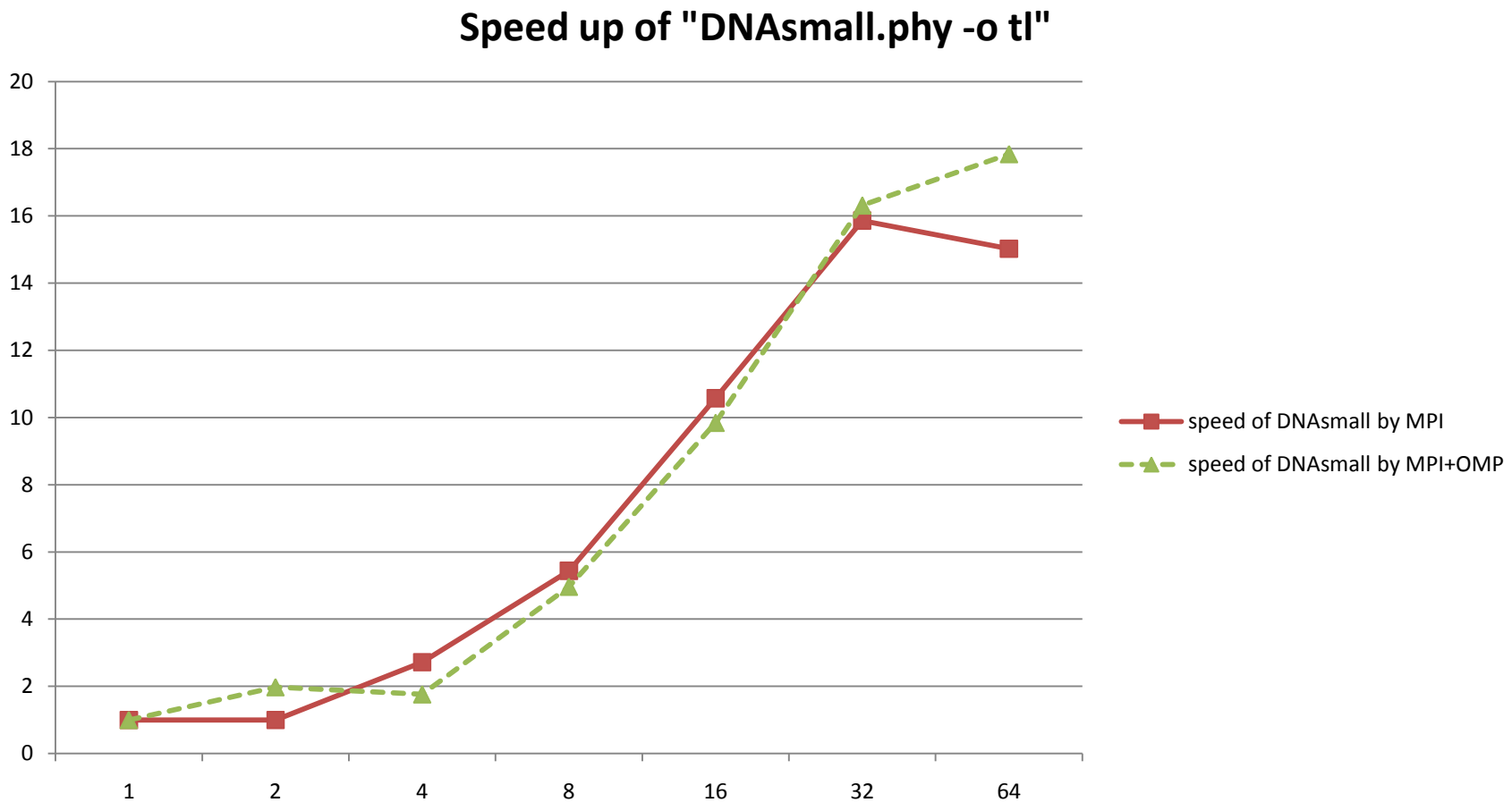
- Master node: distribute *n* jobs to each slave node
- Slave node:Compute with 2 threads (This is done by OpenMp)
- How many jobs should be run with one-time OpenMP?
- Method1



Method2

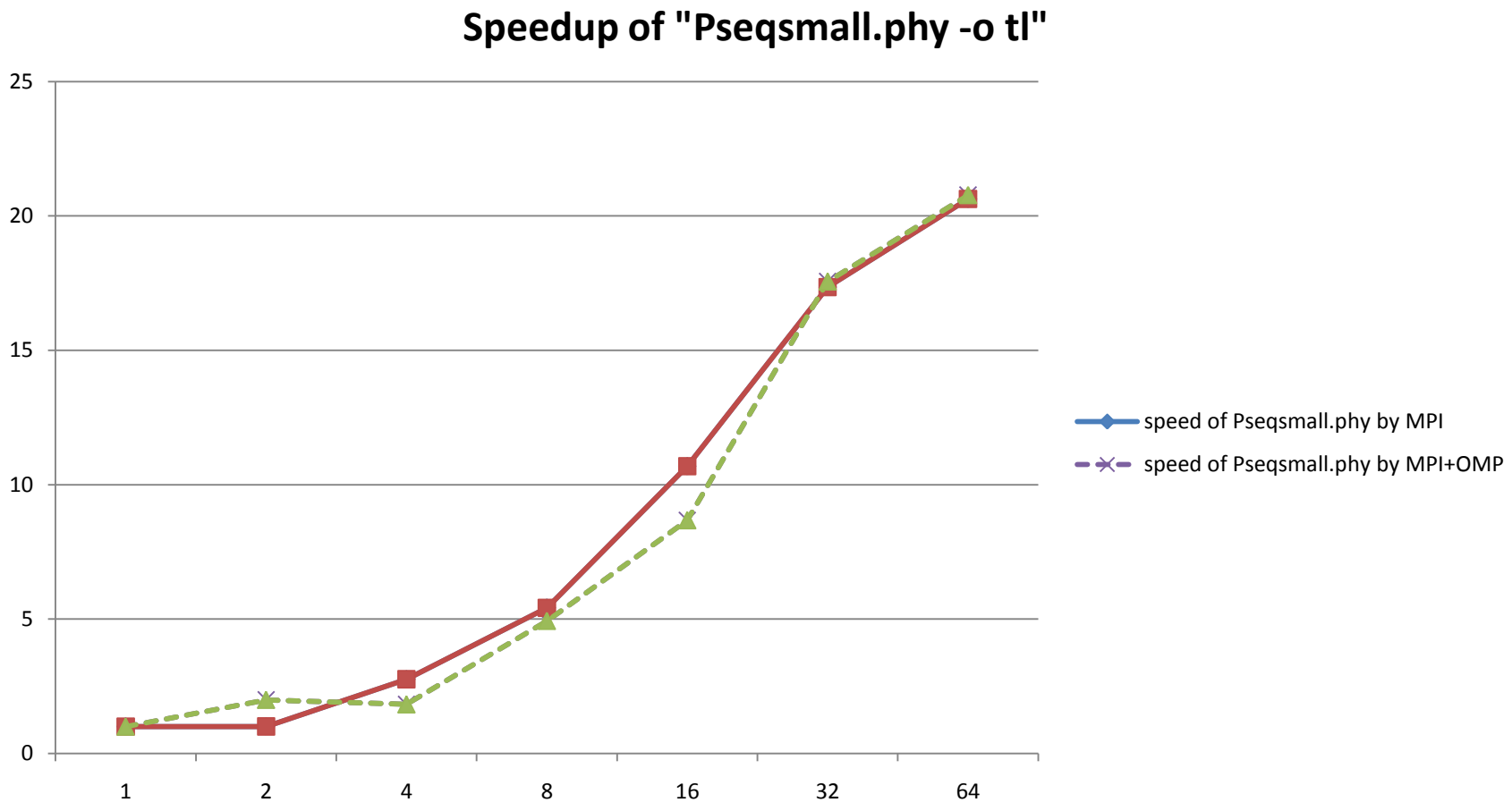
DNAsmall.phy

parallelpc4-> 16 node cluster		MPI	MPI	MPI	MPI	MPI	MPI
np	1	2	4	8	16	32	64
DNAsmall.phy -o tl	9m31s	9m31s	3m30s	1m45s	54s	36s	38s
seconds	571	571	210	105	54	36	38
speed of DNAsmall by MPI	1	1	2.71904761	5.43809523	10.5740740	15.8611111	15.02632
	serial	OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP
np	1	2	2	4	8	16	32
DNAsmall.phy -o tl	9m31s	4m49s	5m24s	1m55s	58s	35s	32s
seconds	571	289	324	115	58	35	32
speed of DNAsmall by MPI+OMP	1	1.97577855	1.76234567	4.96521739	9.84482758	16.3142857	17.84375


Speed up of "DNAsmall.phy -o tl"
20
18
16
14
12
10
8
6
4
2
0
1
2
4
8
16
32
64
speed of DNAsmall by MPI
speed of DNAsmall by MPI+OMP

Pseqsmall.phy

parallelpc4-> 16 node cluster		MPI	MPI	MPI	MPI	MPI	MPI
np	1	2	4	8	16	32	64
Pseqsmall.phy -o tl	24m35s	24m25s	8m54s	4m32s	2m18s	1m25s	1m11s
seconds	1475	1465	534	272	138	85	71
speed of Pseqsmall.phy by MPI	1	1.00682594	2.762172285	5.422794118	10.6884058	17.35294118	20.63380282
	serial	OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP	MPI+OpenMP
np	1	2	2	4	8	16	32
Pseqsmall.phy -o tl	24m35s	12m19s	13m27s	4m59s	2m50s	1m24s	1m11s
seconds	1475	739	807	299	170	84	71
speed of Pseqsmall.phy by MPI+OMP	1	1.99594046	1.827757125	4.933110368	8.676470588	17.55952381	20.77464789


Speedup of "Pseqsmall.phy -o tl"
25
20
15
10
5
0
1
2
4
8
16
32
64
speed of Pseqsmall.phy by MPI
speed of Pseqsmall.phy by MPI+OMP

Observation between MPI and MPI+OpenMP

- MPI occupies one node for distributing the jobs to slave nodes.
- MPI spends more time on system mode than user mode.
- MPI doesn’t have good speedup as OpenMP.