Acceleration of large-scale phylogenetic analyses with non-homogeneous substitution models: implementation and performance evaluation on T2K-Tsukuba super-cluster system

Sohta Ishikawa

University of Tsukuba, Graduate School of Life and Environmental Sciences
University of Tsukuba, Graduate School of Systems and Information

Interdisciplinary Computational Science Program in CCS
April 2012 – March 2013 (REALPHYL)

T2K-Tsukuba General Use Program
April 2013 – March 2014 (NONHOMO)

Outline

Purpose

- Acceleration of large-scale maximum-likelihood phylogenetic analyses with Non-Homogeneous substitution models

Materials & Methods

- “NHML” and “GG98 model”
- MPI/OpenMP parallelization for likelihood calculation algorithm
- Parallel likelihood calculation of multiple trees (tree searching)

Results

- Analyses with simulated sequence data sets (~130 taxa, ~1,0000 nt)
- Good performance for HYBRID parallelization regardless of datasize
- More than 400 times speeding-up in the use of 1,024 CPU cores

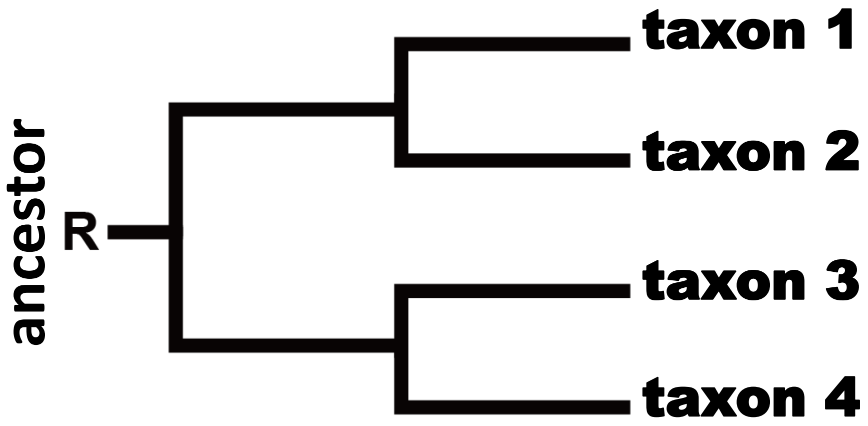
Phylogenetic Analyses

Nucleotide or Amino-Acid Sequences → **Phylogenetic Trees**

Sites

Taxa	
taxon 1	CTTGGCTGTGAACA
taxon 2	GAATAATGTGTAGA
taxon 3	CAACACTCTGGGTA
taxon 4	GCATACTGTGCCGA

N (taxa) × *M* (sites) matrix

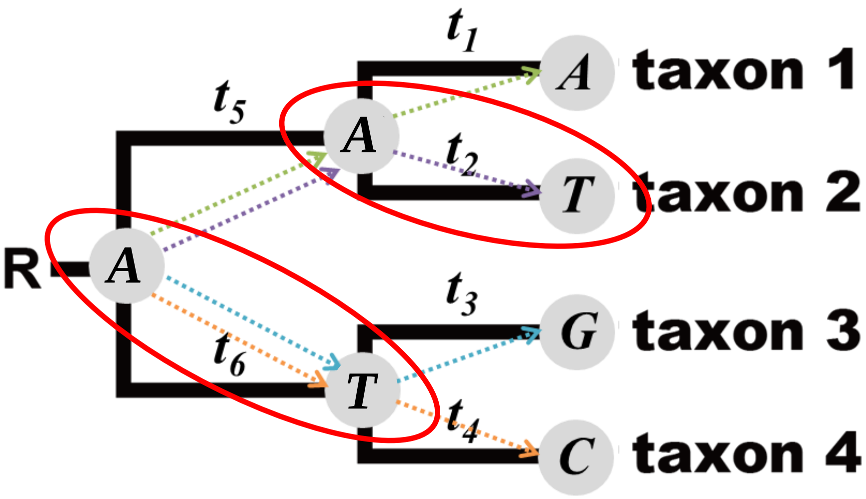


Substitution Models

to

A	C	G	T	
$-r_A$	q_{AC}	q_{AG}	q_{AT}	A
q_{CA}	$-r_C$	q_{CG}	q_{CT}	C
q_{GA}	q_{GC}	$-r_G$	q_{GT}	G
q_{TA}	q_{TC}	q_{TG}	$-r_T$	T

from



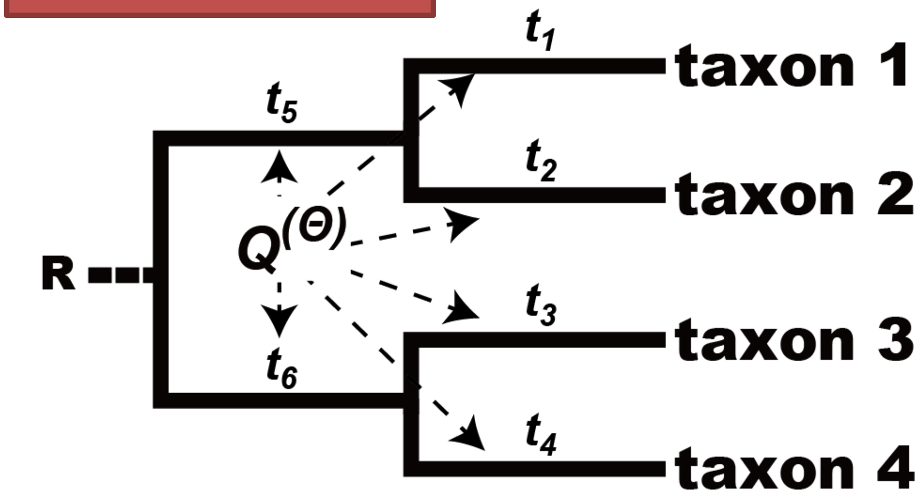
q_{ij} = instantaneous rate for the substitution from i to j

Branch Length (t) = the expected numbers of substitution per site

independent substitution events between branches → same rate ?

Homogeneous and Non-Homogeneous Models

Homogeneous

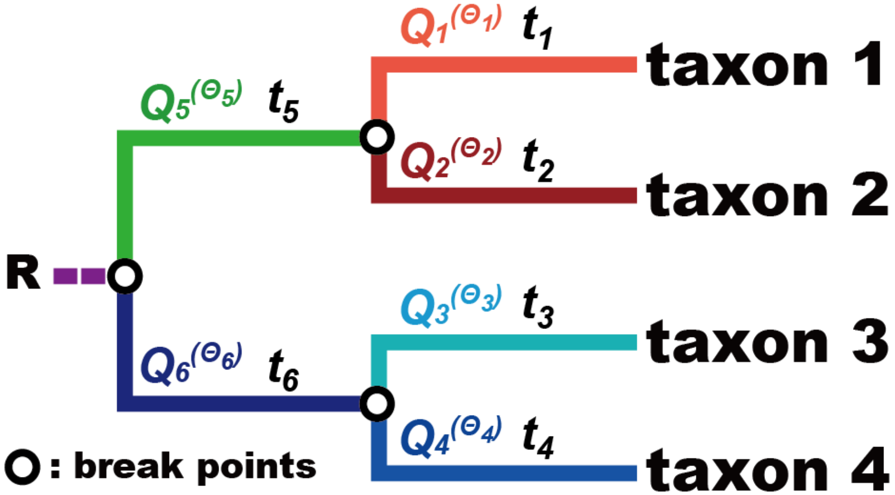


Single model $Q^{(\Theta)}$ to all branches

Assumption

All sequences should be evolved following same substitution process

Non-Homogeneous



Different models $Q_1^{(\Theta 1)}$ ~ $Q_6^{(\Theta 6)}$ to each branch

Assumption

Each sequence can be evolved following independent processes

NH Models : Importance and Problem

- Overcoming the phylogenetic artifacts caused by the heterogeneity of evolutionary processes (e.g, compositional bias)
- Accurate inference of phylogenetic relationships among diverse organisms

computationally intense

For ML tree searching, *N* taxa, *M* sites

Homogeneous Models : $\boldsymbol{O}(\boldsymbol{N}^{3} \times \boldsymbol{M})$

NH Models : $\boldsymbol{O}(\boldsymbol{N}^{4} \times \boldsymbol{M})$

N times longer time is needed for NH

- Computational cost limits the application of NH models into analyses with large-scale real-world sequence datasets

Program must be Parallelized !

- Using large-scale computing systems (super-cluster) for large-scale phylogenetic analyses has been easy, however …

T2K-Tsukuba

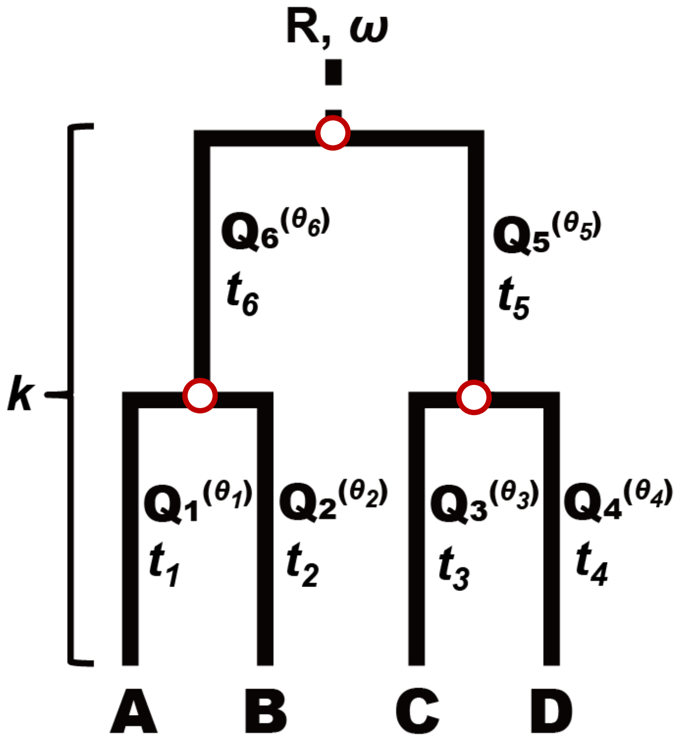
HA-PACS

No parallelized programs for NH models

Purpose

- Large-scale ML phylogenetic analyses with NH models on super-cluster systems

Target : NHML and GG98 model



ω: ancestral G+C content

θ: frequency of G+C occurrence

k: Ts / Tv ratio ○ : break point

Substitution matrix Q

to

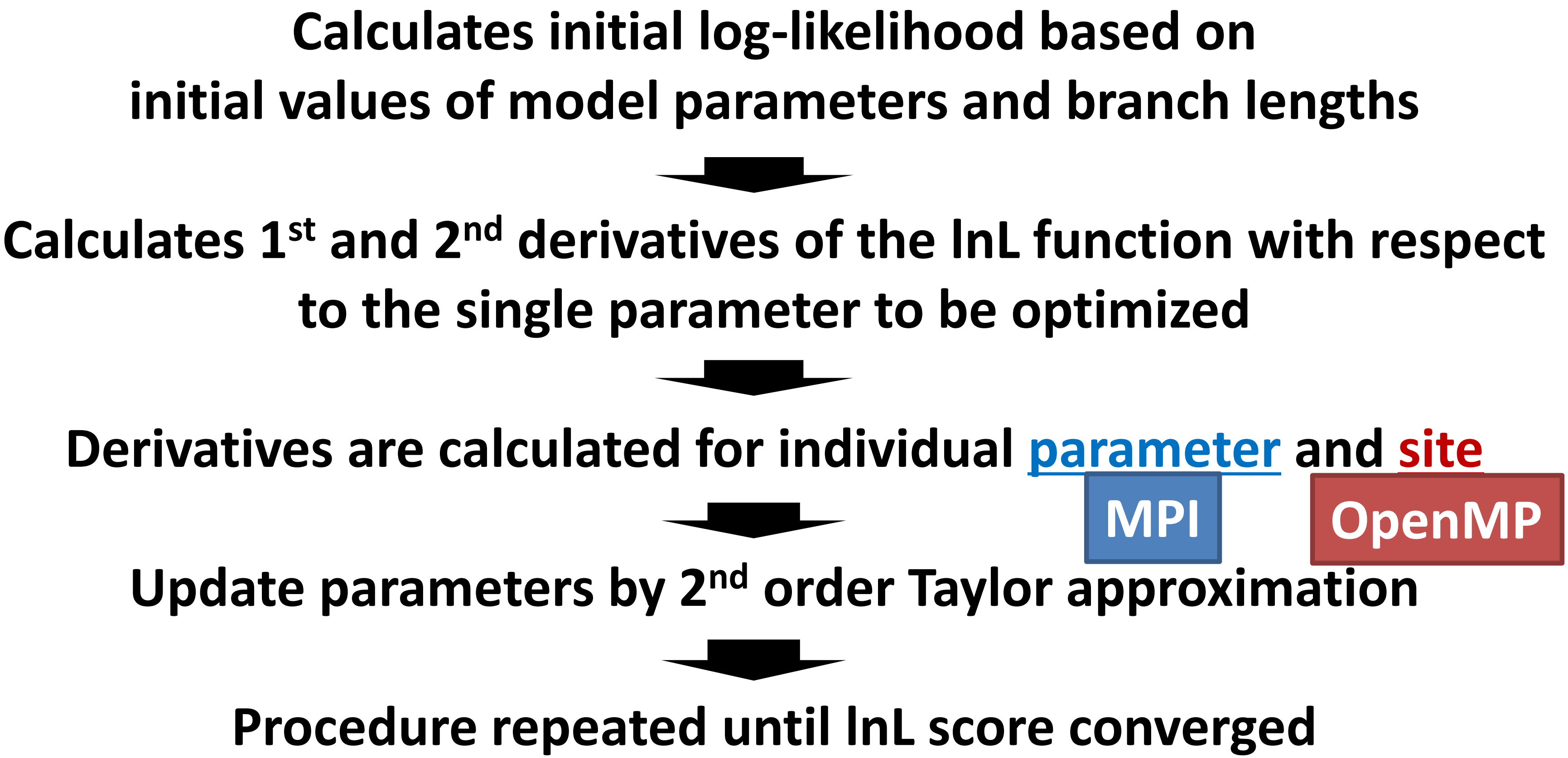
$$
\begin{array}{cccc|c}
A & C & G & T & \\
\hline
-r_A & \frac{\theta}{2} & \kappa\cdot\frac{\theta}{2} & \frac{1-\theta}{2} & A \\
\frac{1-\theta}{2} & -r_C & \frac{1-\theta}{2} & \kappa\cdot\frac{1-\theta}{2} & C \\
\kappa\cdot\frac{1-\theta}{2} & \frac{\theta}{2} & -r_G & \frac{1-\theta}{2} & G \\
\frac{1-\theta}{2} & \kappa\cdot\frac{\theta}{2} & \frac{\theta}{2} & -r_T & T
\end{array}
$$

from

（GG98 model : Galtier and Gouy. *Mol. Biol. Evol* 1998）

Can take heterogeneity of G+C content across a tree into account

lnL Calculation by Newton-Raphson method



Calculates initial log-likelihood based on initial values of model parameters and branch lengths

Calculates 1st and 2nd derivatives of the lnL function with respect to the single parameter to be optimized

Derivatives are calculated for individual parameter and site

Update parameters by 2nd order Taylor approximation

Procedure repeated until lnL score converged

- ➢ Occupies more than 90% in total execution time
- ➢ Calculation of derivatives for each parameter and each site can be independently operated

Dataset and Environment

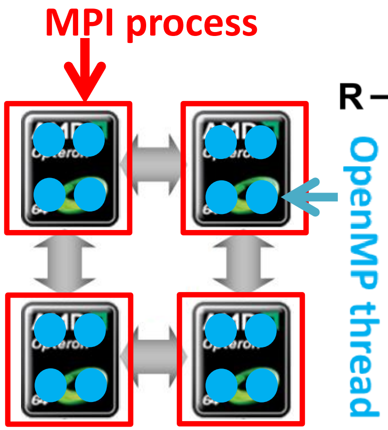
Dataset

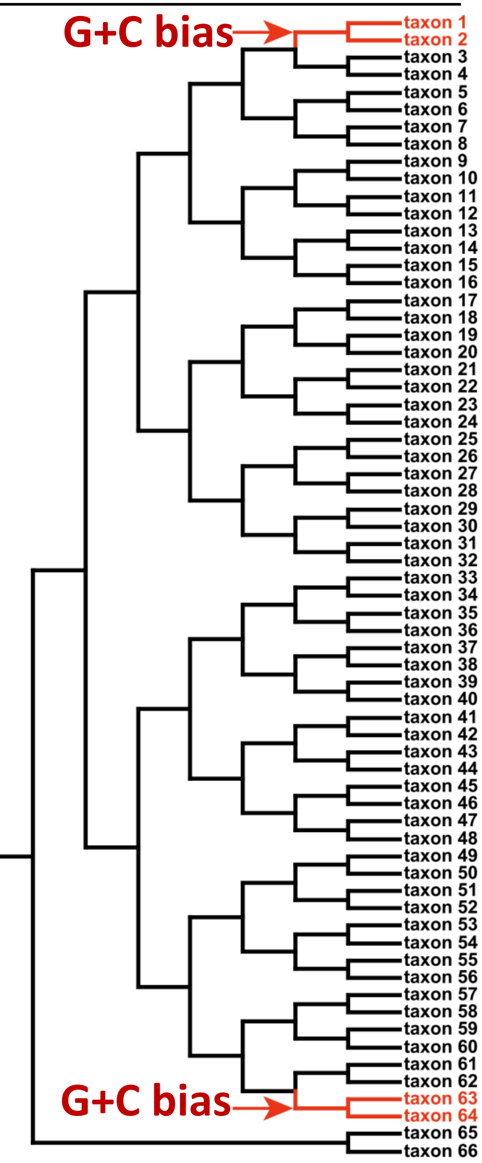
- Simulated nt sequences based on the model tree
- 66 taxa and 130 taxa model trees
- 66 taxa, 2,500/10,000 sites & 130 taxa, 2,500 sites data sets
- 24 (66 taxa) and 48 (130 taxa) alternative trees including “true” tree
- Measured total execution time for lnL calculation of all topologies

Environment

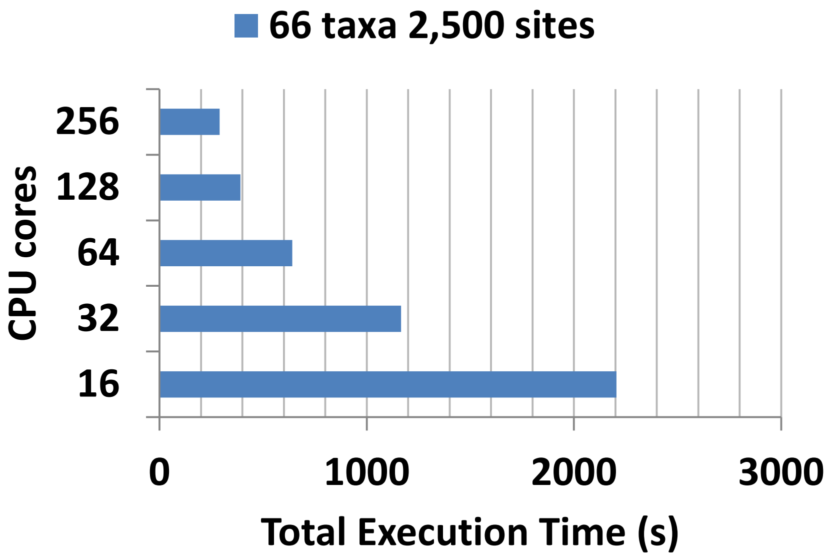
T2K Tsukuba super cluster

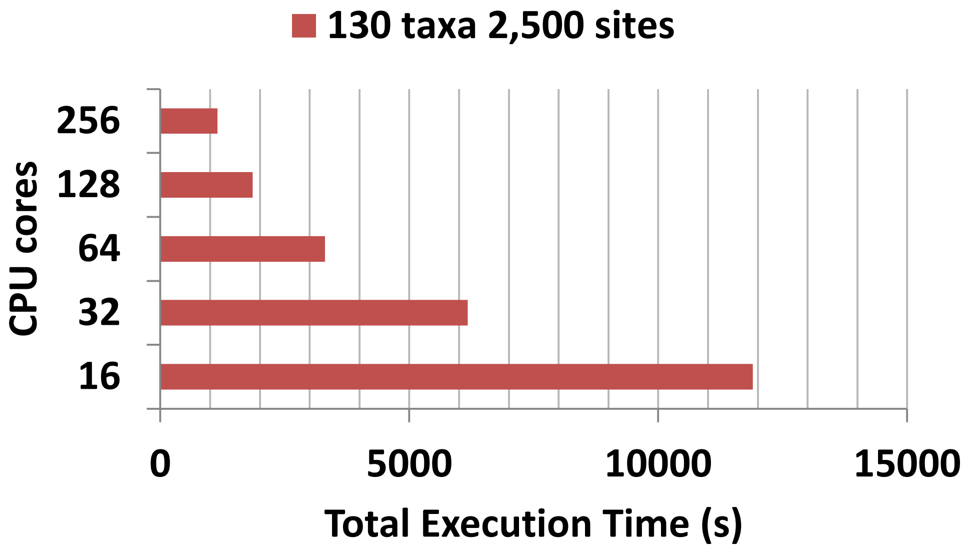
CPU	Quad-core AMD Opteron 8356 (2.30GHz) （4 cores×4 sockets / node）
Memory	2 * 16 GB DDR2 667MHz / node
Network	Infiniband DDR x 4 rail
Compiler	GCC 4. 6. 4
MPI Library	MVAPICH2 Ver. 1. 7

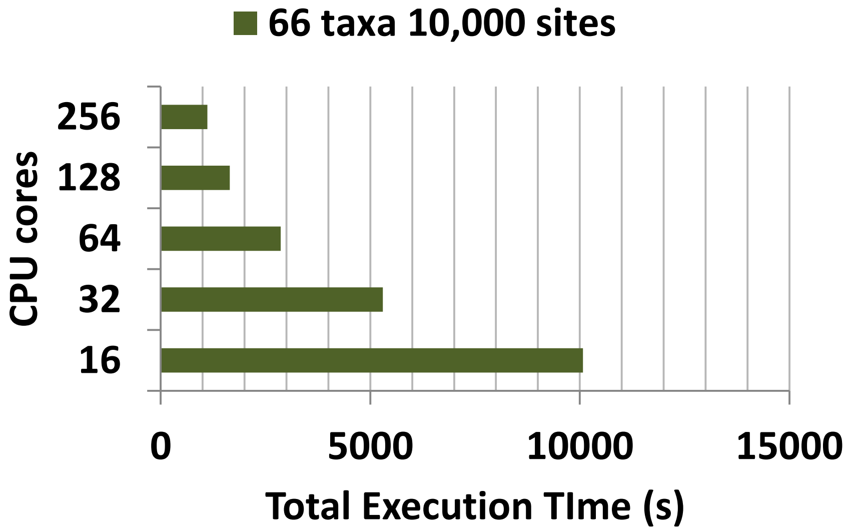




Speeding-Up in Three Data Analyses

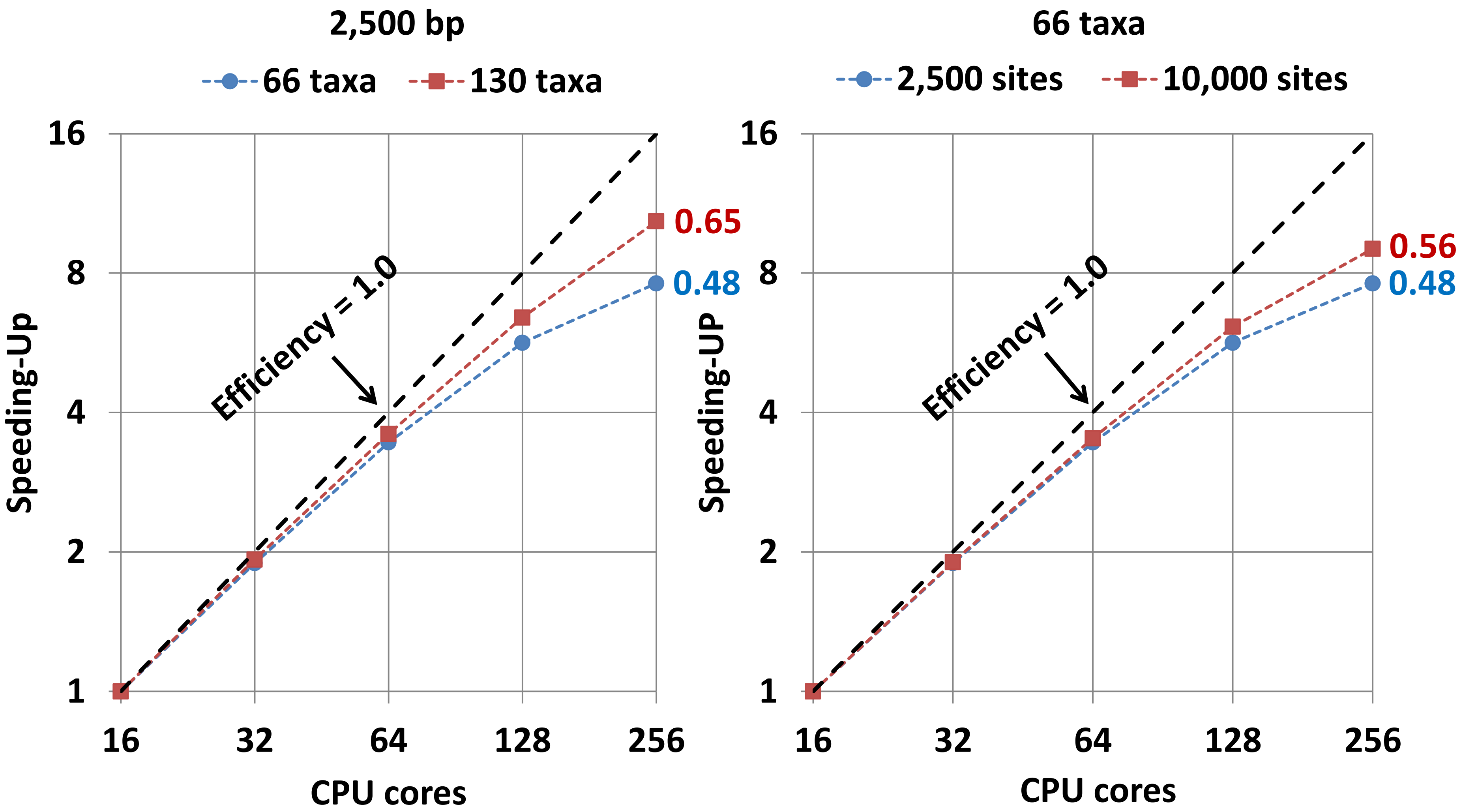






* True tree was successfully selected as ML tree from all data analyses

Parallel Efficiency: Different Number of Taxa and Sites



Pros and Cons of HYBRID Parallelization for NR

- Good performance regardless of datasize (taxa and sites)

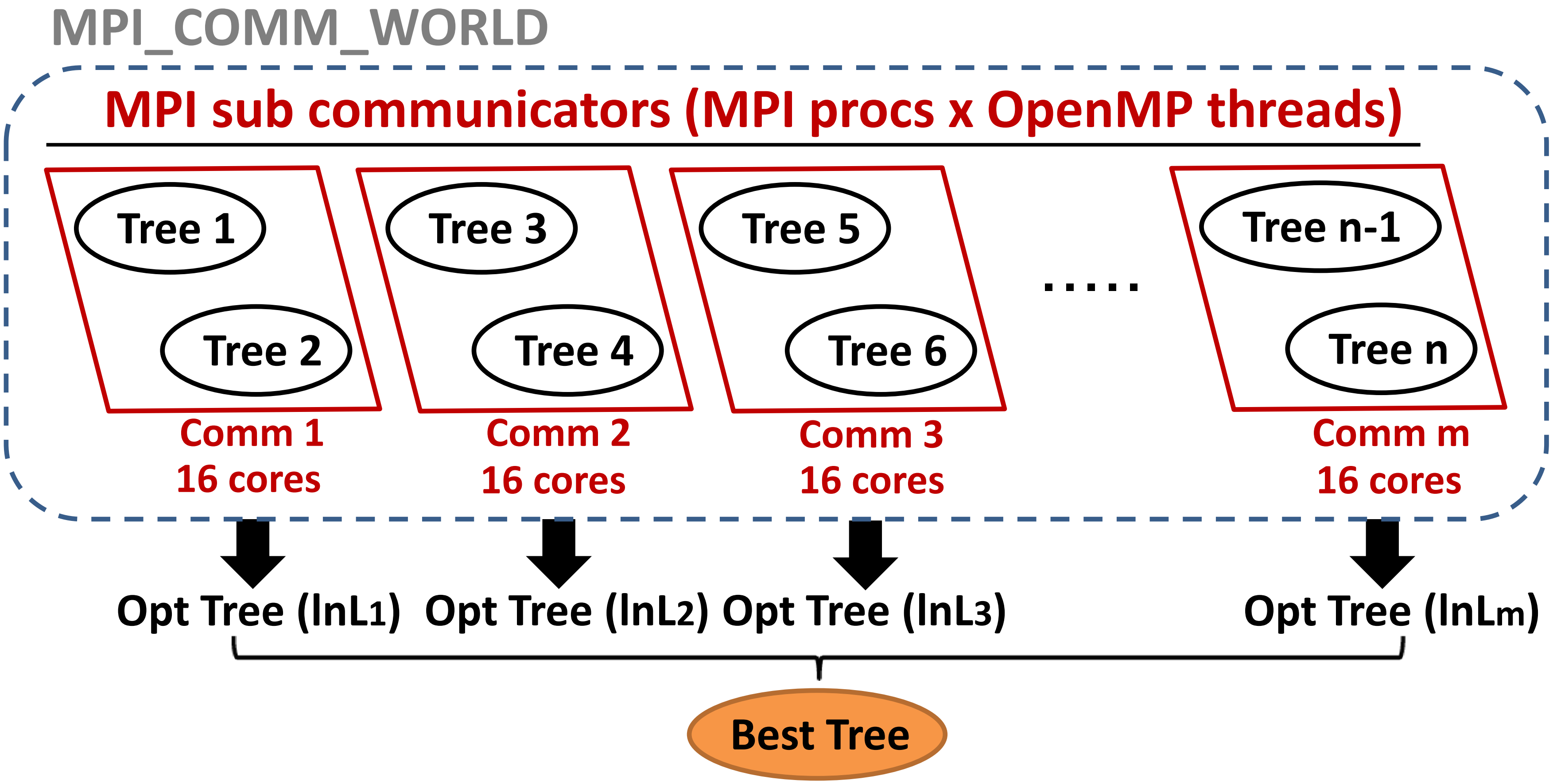
- Efficiency decreased as number of CPU cores increased
  - ✓ Increment of datasize and time for MPI communication

Performance limit on the parallelization for lnL calculation of the single tree

Parallel lnL Calculation for Multiple Trees

Serial lnL calculation for multiple trees

Parallel lnL calculation



MPI_COMM_WORLD

MPI sub communicators (MPI procs x OpenMP threads)

Tree 1 Tree 2

Comm 1
16 cores

Tree 3 Tree 4

Comm 2
16 cores

Tree 5 Tree 6

Comm 3
16 cores

.....

Tree n-1 Tree n

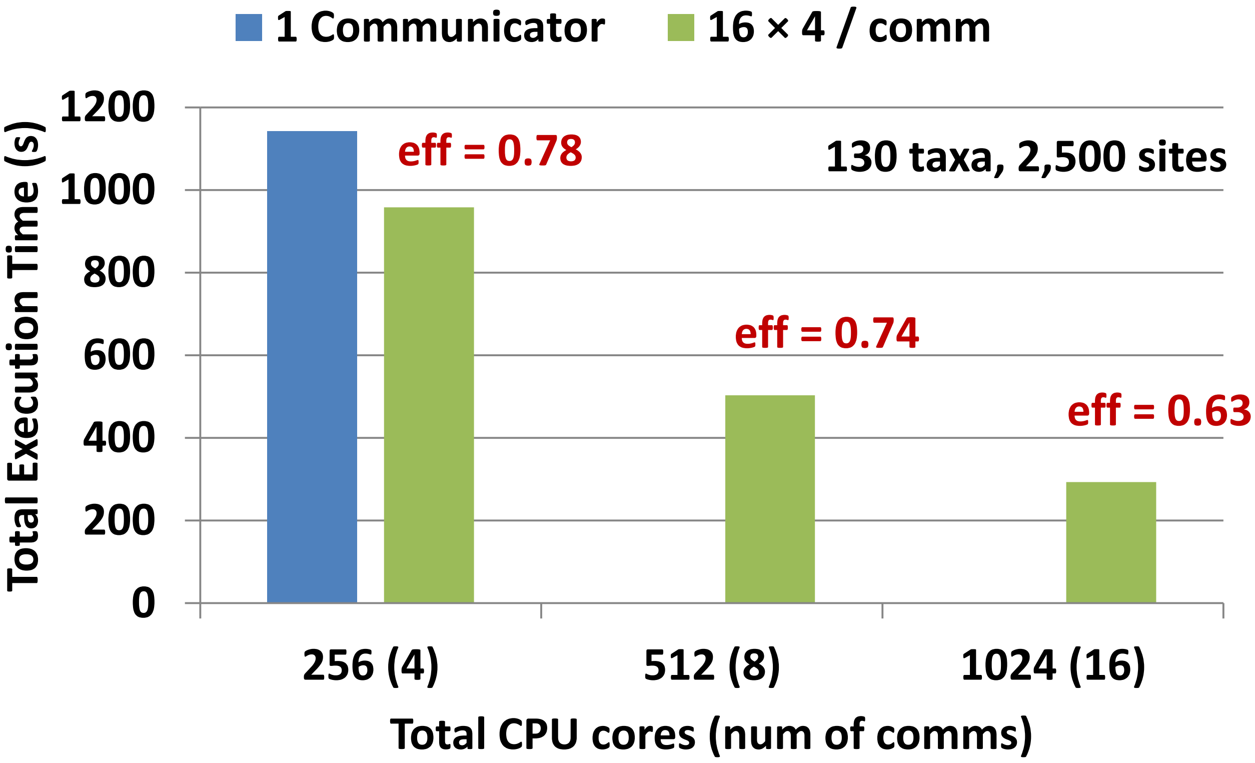
Comm m
16 cores

Opt Tree (lnL$_1$) Opt Tree (lnL$_2$) Opt Tree (lnL$_3$) Opt Tree (lnL$_m$)

Best Tree

Further Speeding-Up by Parallel lnL Calculations

- 64 CPU cores (4 nodes, 16 MPI processes × 4 OpenMP threads) per sub communicator
- 256, 512, 1,024 CPU cores (number of comms was changed)
- lnLs of 48 alternative trees were calculated in parallel



Contribution of Parallel lnL Calculation

- 40 times speeding-up with 1,024 CPUs than 16 CPUs
  - ✓ 400 times speeding-up compared with serial (1 CPU) version

- Good efficiency (> 0.6) with more than 1,000 total CPU cores

- Flexible parallelism (number and size of communicators) for various data analyses

- Application into Maximum-likelihood tree searching with subtree pruning regrafting (SPR)
  - ✓ 16 hours (256 CPU cores) for the analysis with 30 taxa, 12,500 sites dataset of Marine Cyanobacteria
  - ✓ < 24 hours for 50 taxa data analyses with > 2,000 CPU cores
  - ✓ Check-point function has been already implemented

Conclusion

- MPI/OpenMP HYBRID parallelization of NR method showed good performance regardless of datasize
- Efficient parallel lnL calculation for multiple trees achieved more than 400 times speeding-up with 1,024 CPU cores
- Parallelized NHML is useful application for large-scale real-world data analyses on super-cluster

Publications

1. **Ishikawa et al.** Hybrid MPI/OpenMP parallelization of a phylogenetic program with Non-Homogeneous models: toward the analyses of large-scale sequence datasets. ***High Performance Computing Symposium 2014***
2. **Ishikawa et al.** MPI/OpenMP HYBRID Parallelization for Phylogenetic Analyses based on Non-Homogeneous Substitution Models:Implementation and Performance Evaluation for Large-Scale Computing Systems. ***accepted in IPSJ Transactions on Advanced Computing System. vol. 47***

Future Plans

- **Phylogenomic analyses for global phylogeny of gamma-proteobacteria**
  (Interdisciplinary Computational Science
  Program on COMA PACS IX, Apr 2014 – Mar 2015)

- **Implementation of more flexible Non-Homogeneous substitution models (GTR)**

- **Partial optimization of BLs, LRT**

- **GPU computing, Manycore computing**

Acknowledgement

Related Members

Tetsuo Hashimoto (Univ. Tsukuba, MEM, prof.)
Yuji Inagaki (Univ. Tsukuba, MEM, associate prof.)
Mitsuhisa Sato (Univ. Tsukuba, HPCS, prof.)
Masahiro Nakao (RIKEN AICS, PD)
Miwako Tsuji (RIKEN AICS, PD)

Grants

Japanese Society for the Promotion of Science
(Nos. 21370031 and 22657025)
Ministry of Education, Culture, Sports, Science and Technology
(Nos. 23117006)
Research Fellowships from JSPS for Young Scientists
(Nos. 24007)