第4回「計算科学による新たな知の発見・統合・創出
~PACS-CSシステムと計算科学~」@ April 25, 2008

MOL EVOL
MICROBES
UNIV.TSUKUBA

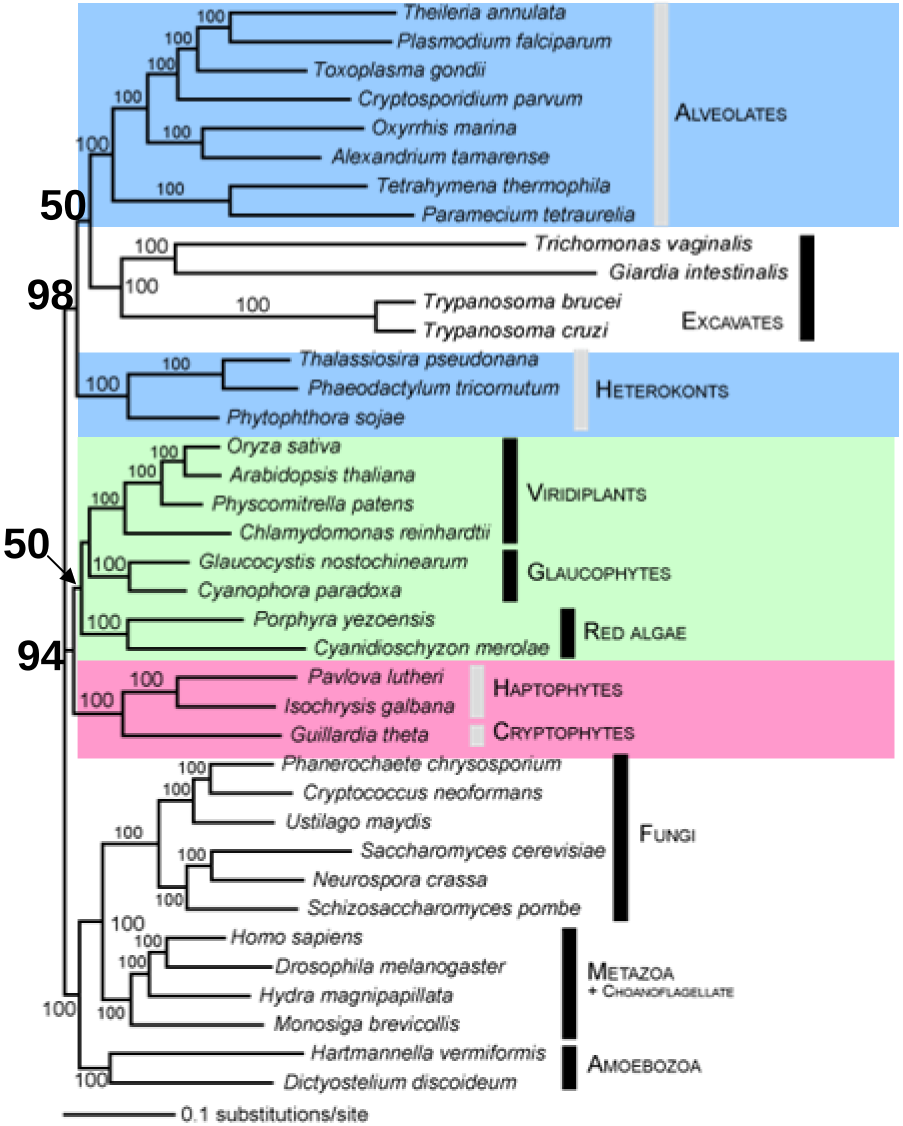
網羅的樹形探索による
ブートストラップ解析法の検討

稲垣祐司
筑波大学
生命環境科学研究科
構造生物学専攻

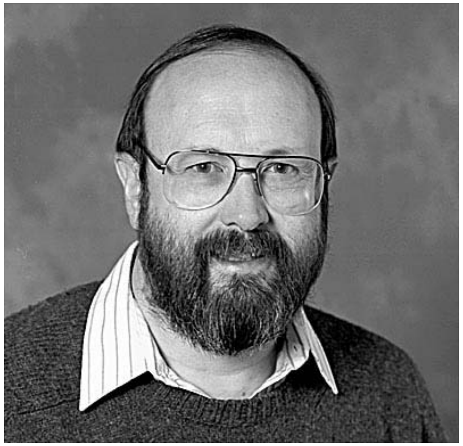
Outline

- How to evaluate the phylogenetic estimate from a real data
  - Bootstrap
- How to search the ML tree
  - Pros & cons in exhaustive and heuristic tree search
- Analyses of a 24-taxon EF-1α data set
  - Impact of the methods for HTS
- ETS considering 2M trees on PACS-CS
  - Evaluate the efficiency of HTS

“Bootstrap” in phylogeny

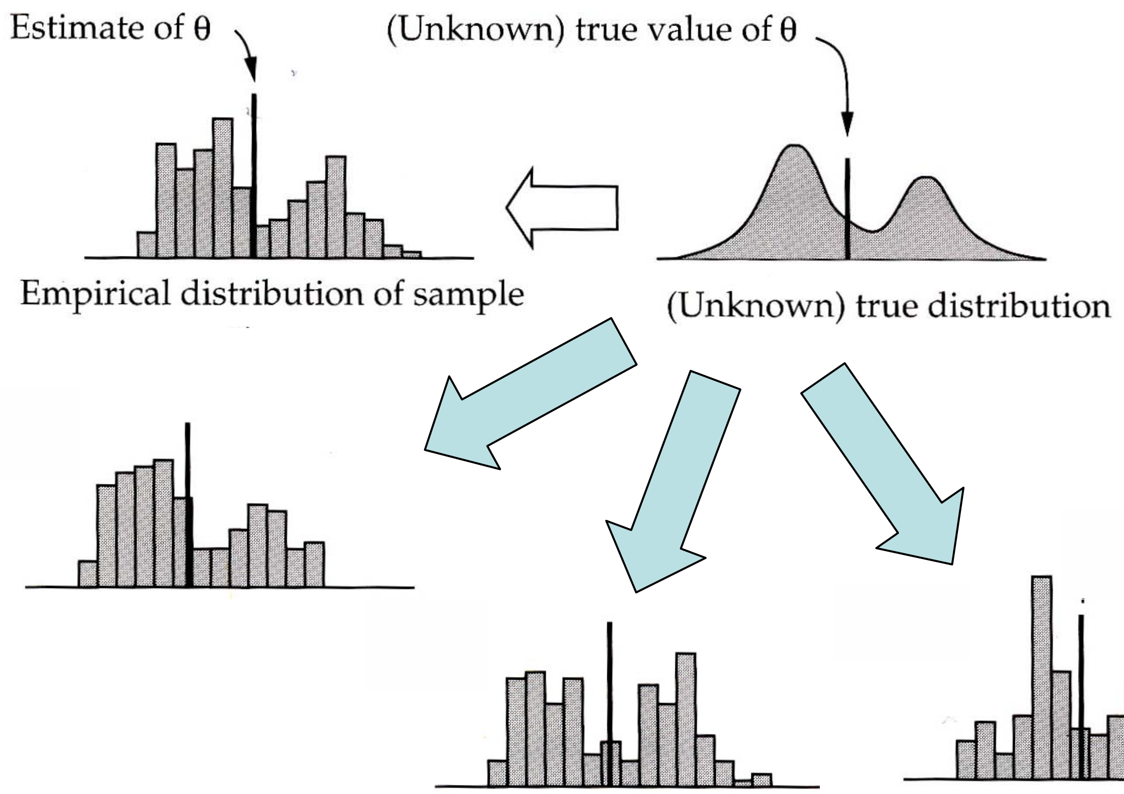


- To evaluate the phylogenetic estimate from a real data
- Can publish no phylogeny without bootstrap values



Joseph Felsenstein

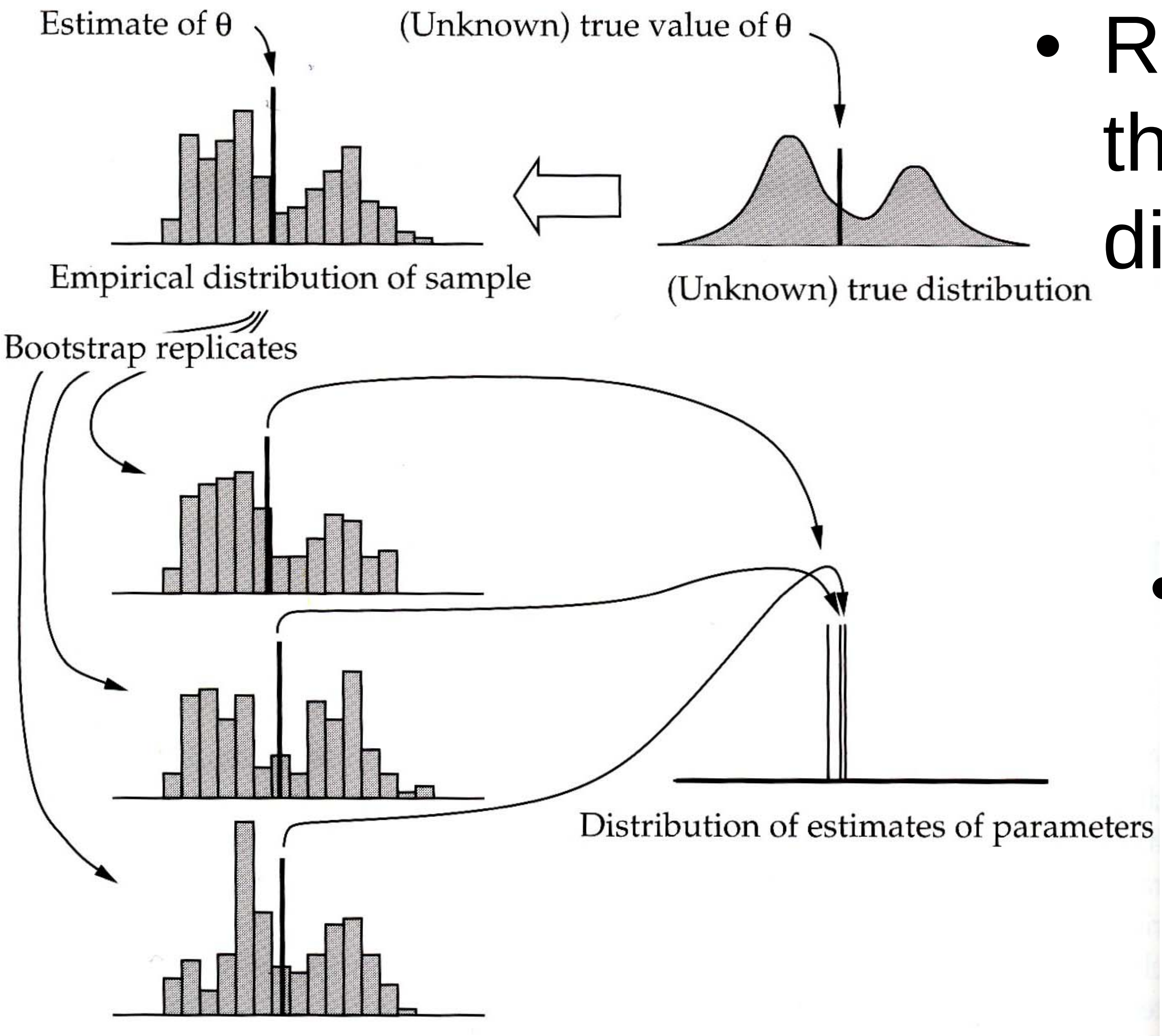
Bootstrap


Estimate of θ
(Unknown) true value of θ
Empirical distribution of sample
(Unknown) true distribution

- Repeat sampling and estimate θ for *N* times

- However, cannot always repeat sampling from the true distribution
- *It's time to BOOTSTRAP!*

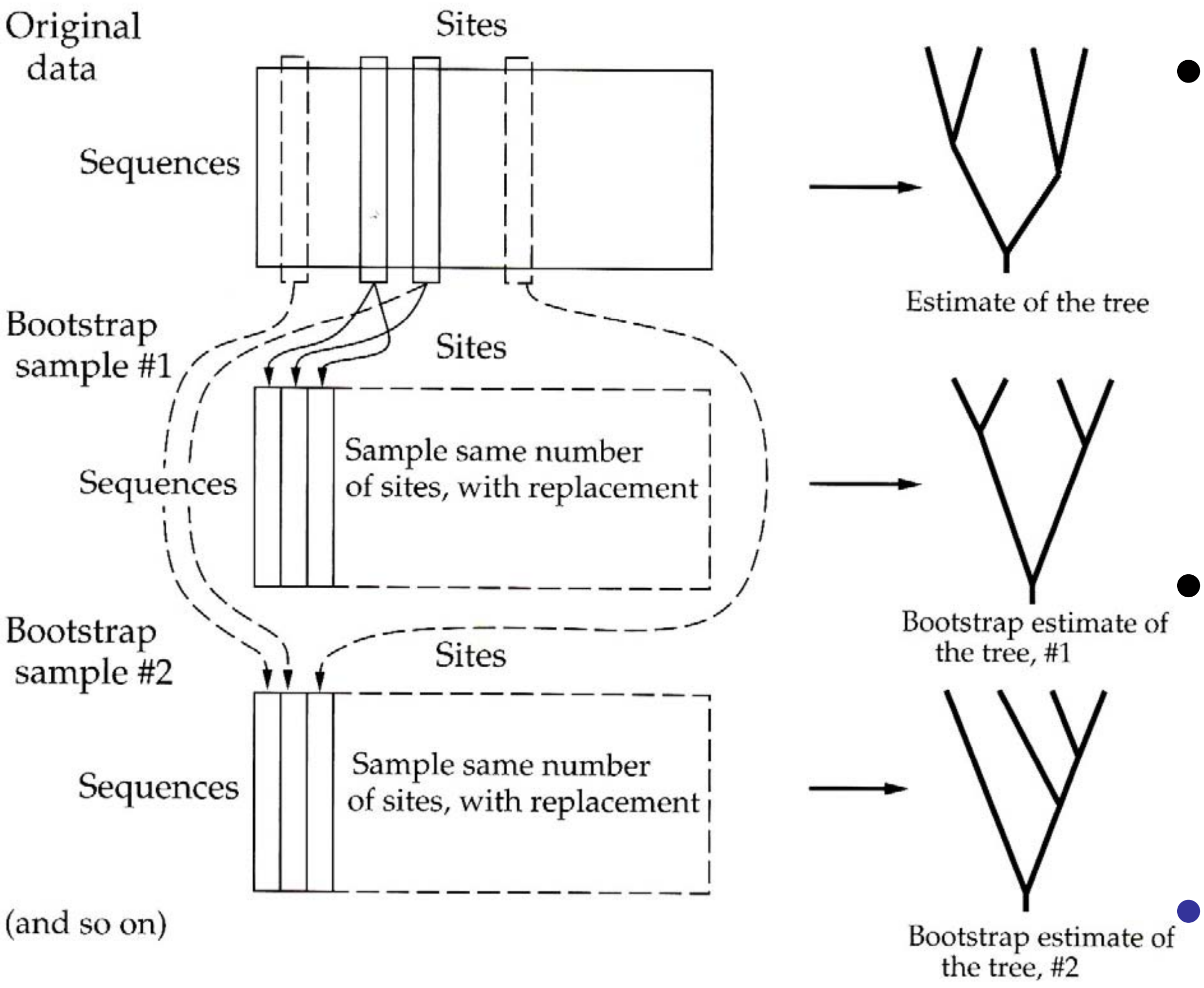
Bootstrap


Estimate of θ
(Unknown) true value of θ
Empirical distribution of sample
(Unknown) true distribution
Bootstrap replicates
Distribution of estimates of parameters

- Resample from the sampled distribution
  - Bootstrap replicates

- Estimate *θ* from bootstrap replicates

Bootstrap in phylogeny


Original data
Sites
Sequences
Estimate of the tree
Bootstrap sample #1
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #1
Bootstrap sample #2
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #2
(and so on)

- BP replicates
  - Same size as the original data
  - *N* ≥100
- Estimate a “BP” tree from each BP replicate
- *Larger is better*

Number of trees to evaluate

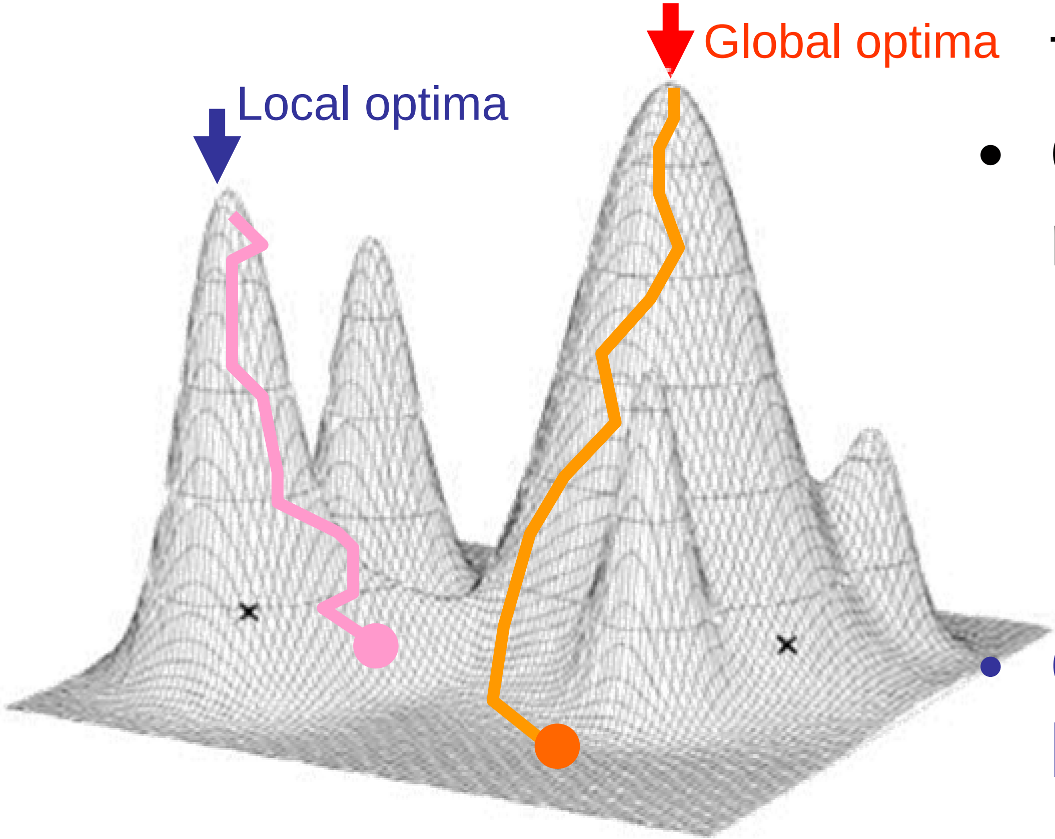
OTU	Number of (unrooted) trees
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
…	…
20	8200794532637891559375

← 2M trees

- The ML tree can be selected by exhaustive tree search (ETS)

Heuristic tree search

- ETS is not always realistic
  - In the real-world data, OTU ≥10
- Heuristic tree search

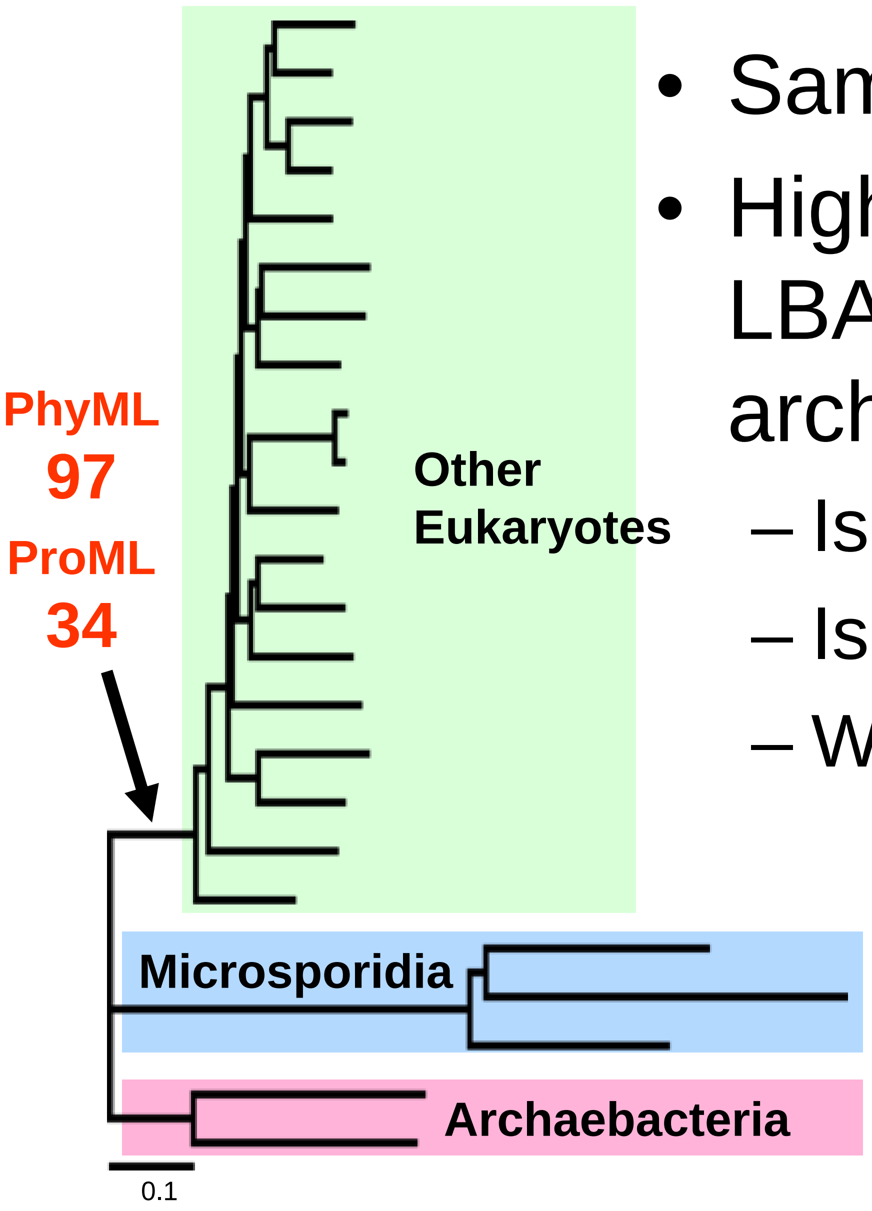


- Not score all possible trees
- Quick & dirty (but realistic)
  - Nearest-neighbor interchanges (NNI)
  - Subtree pruning & regrafting (SPR)
- Can climb up to a local optima

Motivation

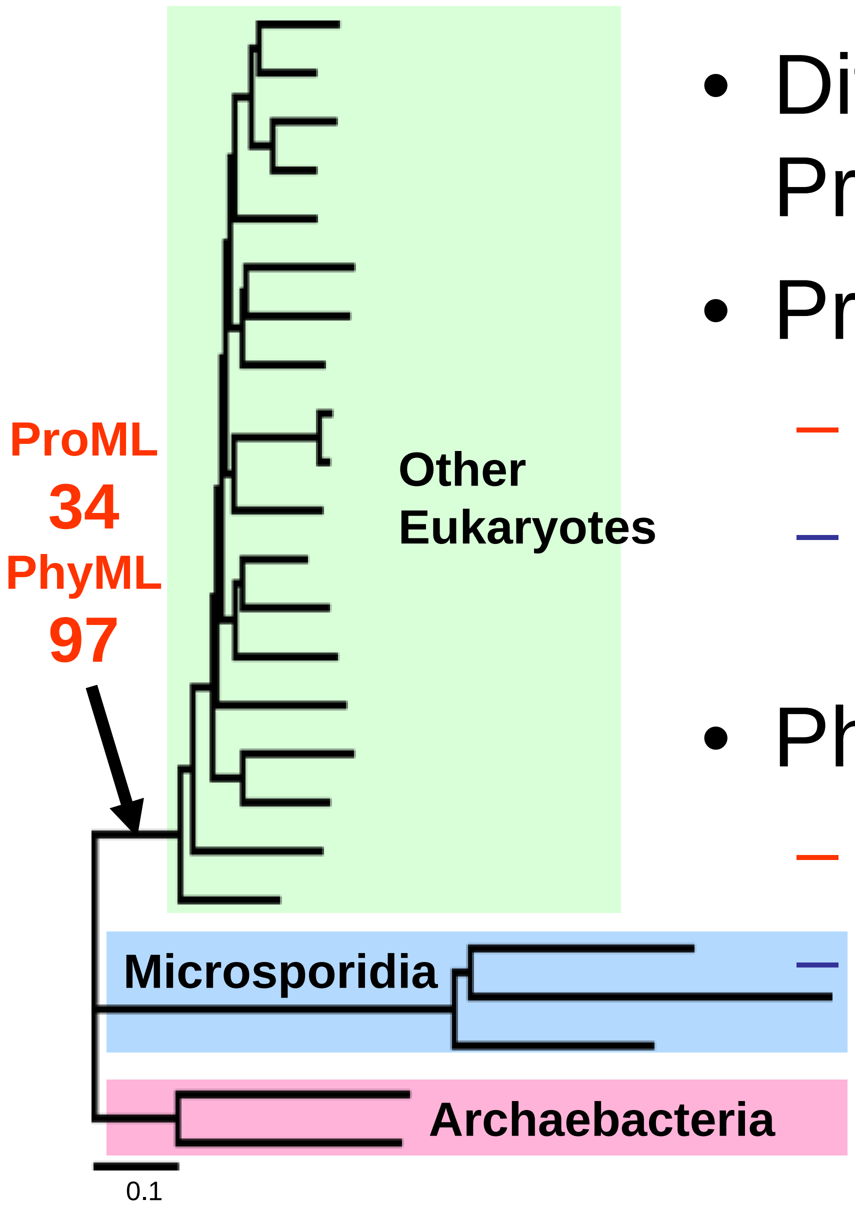
- Data are usually too large for ETS
  - Generally use a HTS to analyze “real-world” data
    - the “ML” tree estimation
    - Bootstrap (Don’t want to repeat ETS for 100 times)
  - Possibly bias the estimate
- Analyses of a 24-taxon EF-1α data set
  - Impact of the methods for HTS
- ETS considering 2M trees on PACS-CS
  - Evaluate the efficiency of HTS

Microsporidian EF-1α



- Same method & model
- High support for M+A is a LBA artifact between archaebac. & microsporidia
  – Is HTS in PhyML sucks?
  – Is HTS in ProML good?
  – Which is closer to the truth?

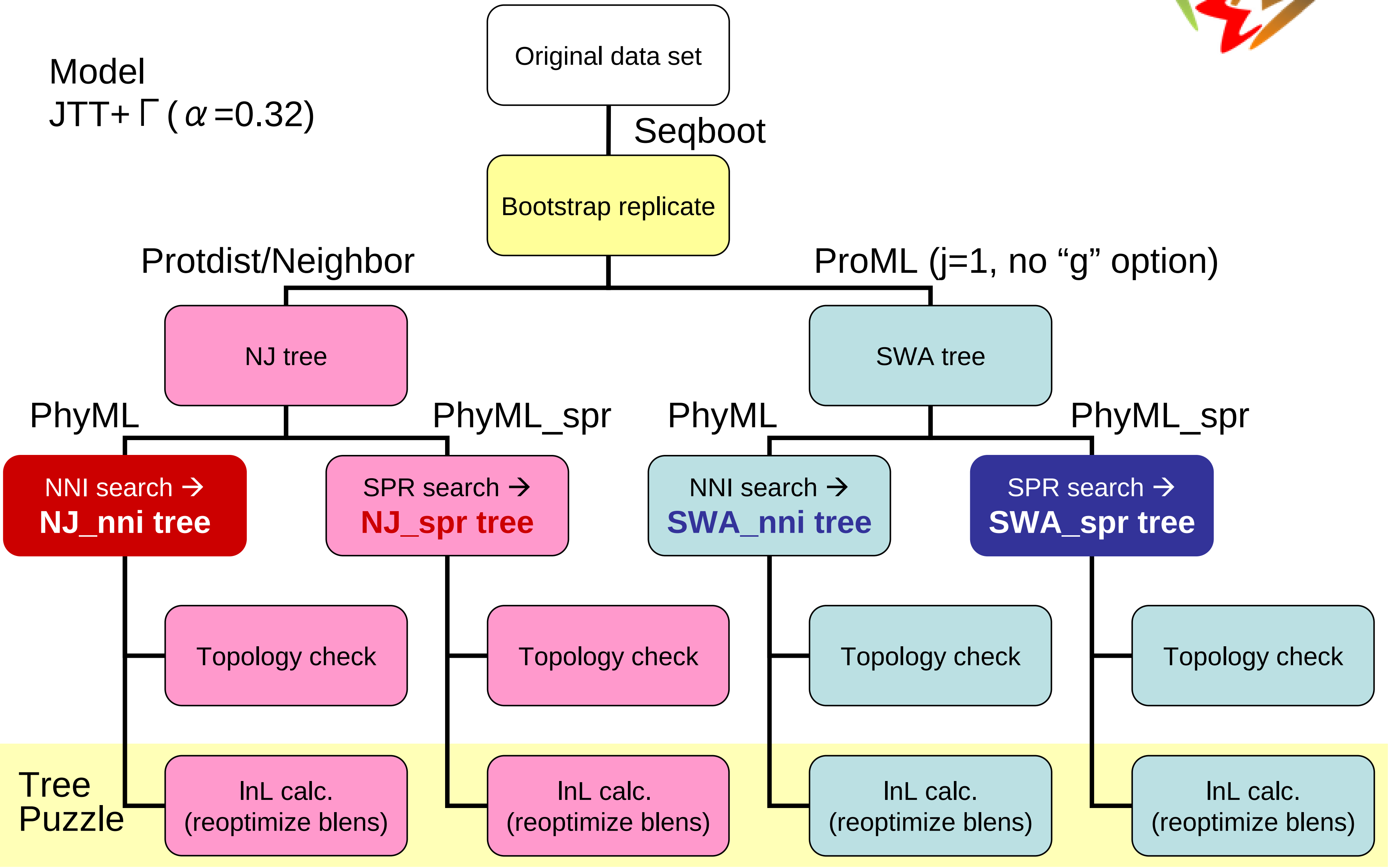
HTS in ProML & PhyML



- Difference in HTS between ProML & PhyML
- ProML
  - – Stepwise addition (SWA)
  - – subtree pruning & regrafting (SPR)
- PhyML
  - – Neighbor-joining (NJ)
  - – nearest neighbor interchange (NNI)

Experimental design

Model
JTT+Γ(α=0.32)



Original data set

Seqboot

Bootstrap replicate

Protdist/Neighbor

ProML (j=1, no “g” option)

NJ tree

SWA tree

PhyML

PhyML_spr

PhyML

PhyML_spr

NNI search →
NJ_nni tree

SPR search →
NJ_spr tree

NNI search →
SWA_nni tree

SPR search →
SWA_spr tree

Topology check

Topology check

Topology check

Topology check

Tree
Puzzle

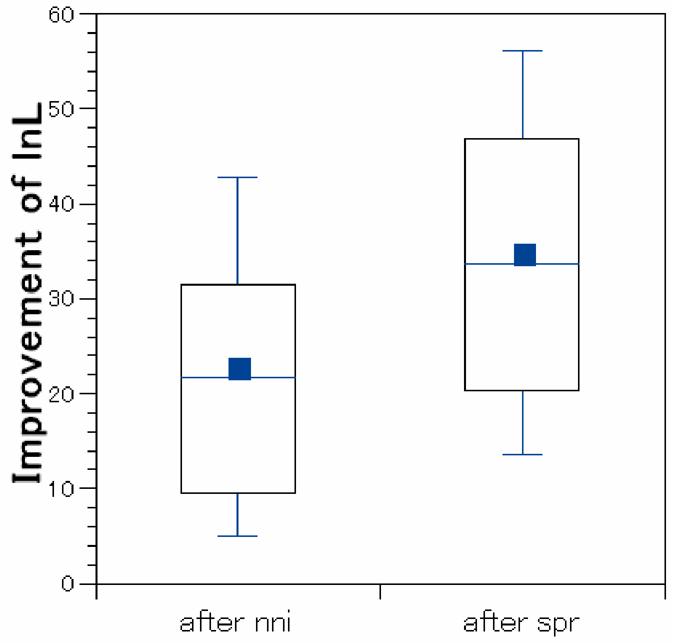
lnL calc.
(reoptimize blens)

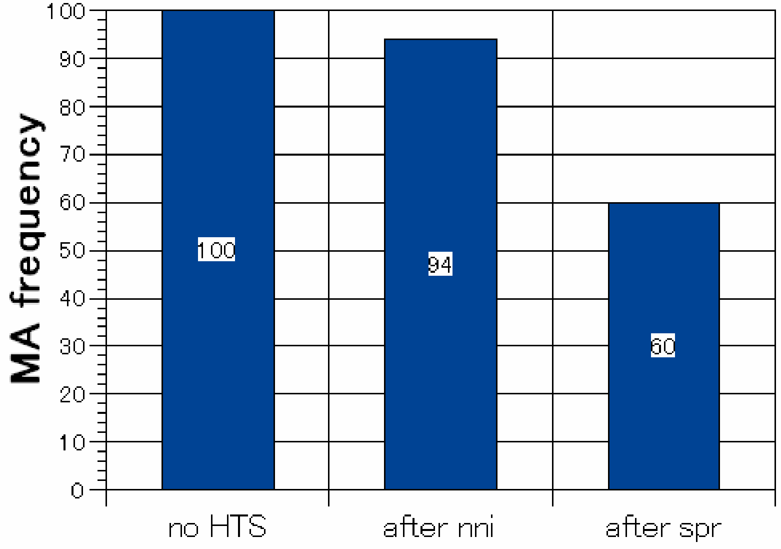
lnL calc.
(reoptimize blens)

lnL calc.
(reoptimize blens)

lnL calc.
(reoptimize blens)

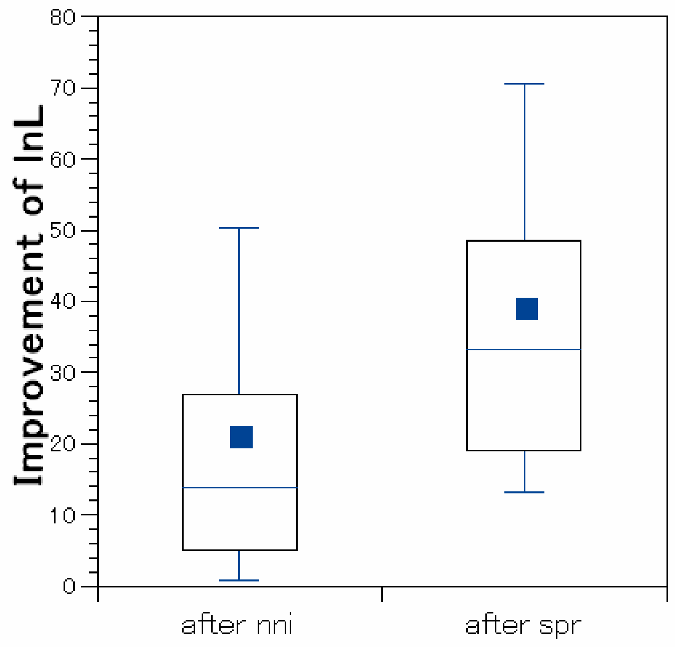
NJ-based searches

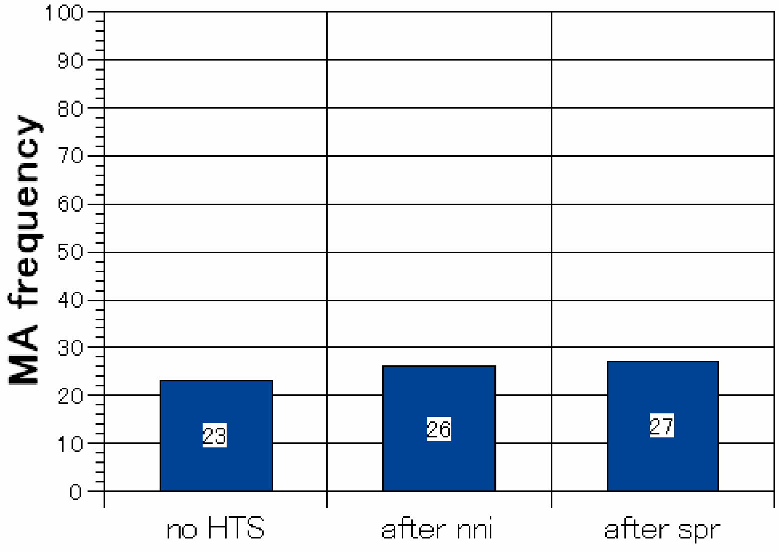




- HTS improved lnL scores; SPR>NNI
- Better HTS, lower the M+A support
  - NJ & NJ_nni (PhyML) overestimated the M+A support

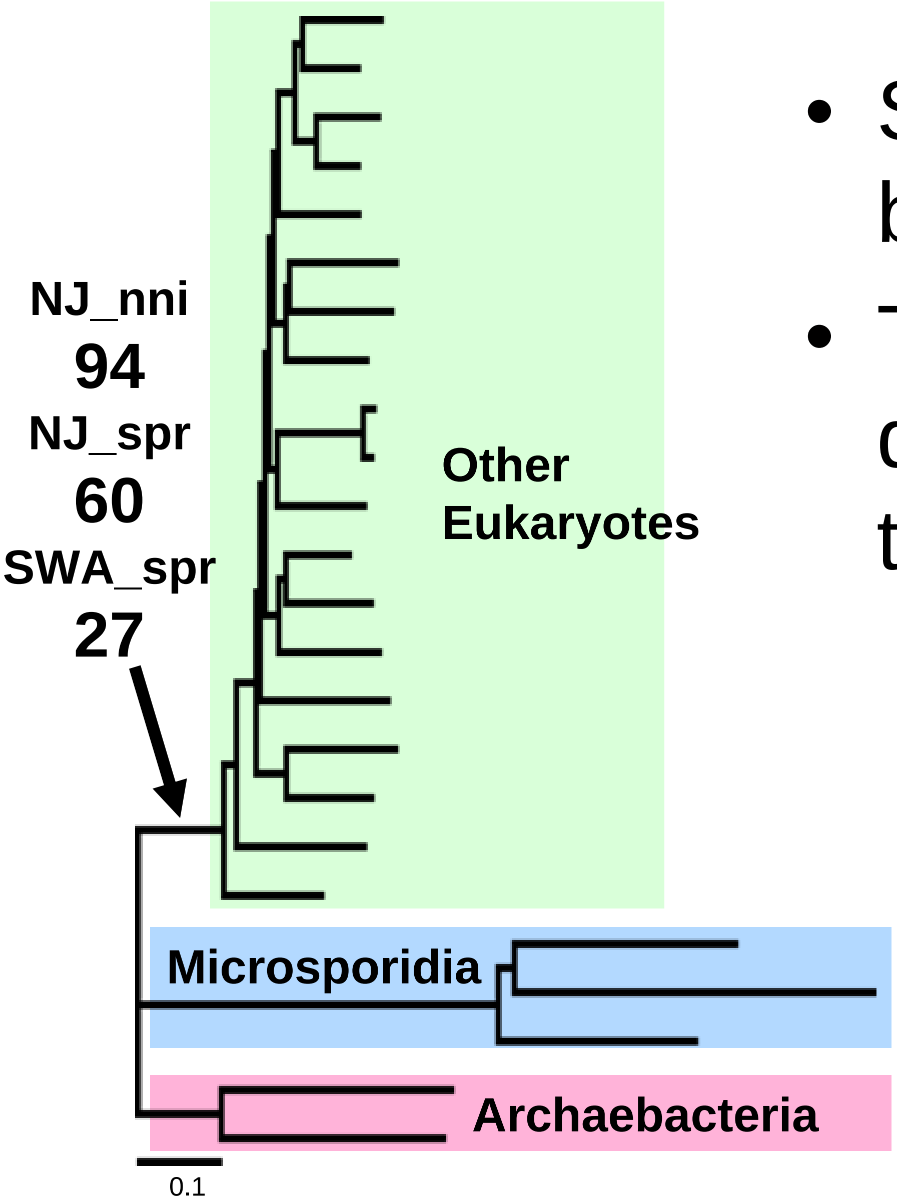
SWA-based searches





- HTS improved the lnL scores
- M+A support didn’t change after HTS
  - – M+A in the initial trees stayed
  - – M-A in the initial trees stayed

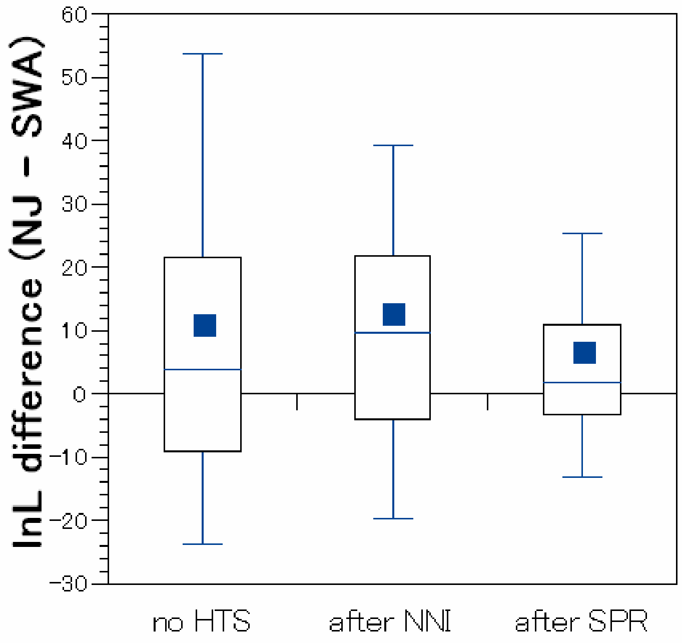
Start tree; NJ or SWA?



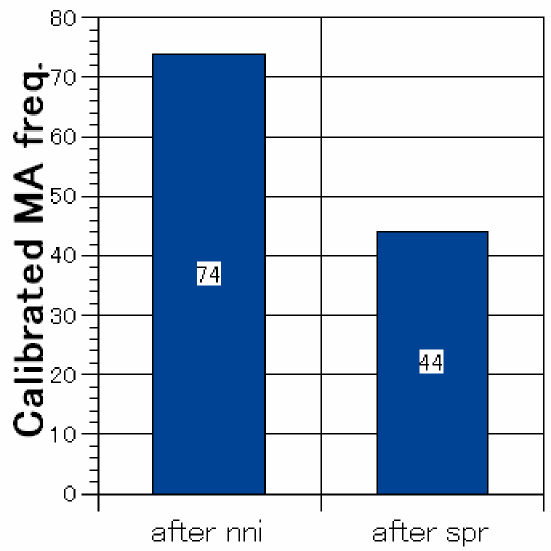
- SPR always found better trees than NNI
- The M+A support depends on the initial trees

  - Need to compare the two results

Start tree; NJ or SWA?



- Neither outperformed the other
  - NJ_spr ≥ SWA_spr
  - The *true* value: somewhere between the two searches?



- “Calibrated” M+A: 44%
  - Closer to the *true* value?
  - NJ_spr overestimated
  - SWA_spr underestimated

Pre PACS-CS analysis

- SPR constantly performs better than NNI
- SPR may search only restricted tree space
  - SWA-based searches tend to separate long branches
  - NJ-based searches tend to put long branches together
  - Two HTS methods migrate toward the same “peak”?
    - *Global* or *local*
- To test the efficiency of HTS, we need to know where is the global peak

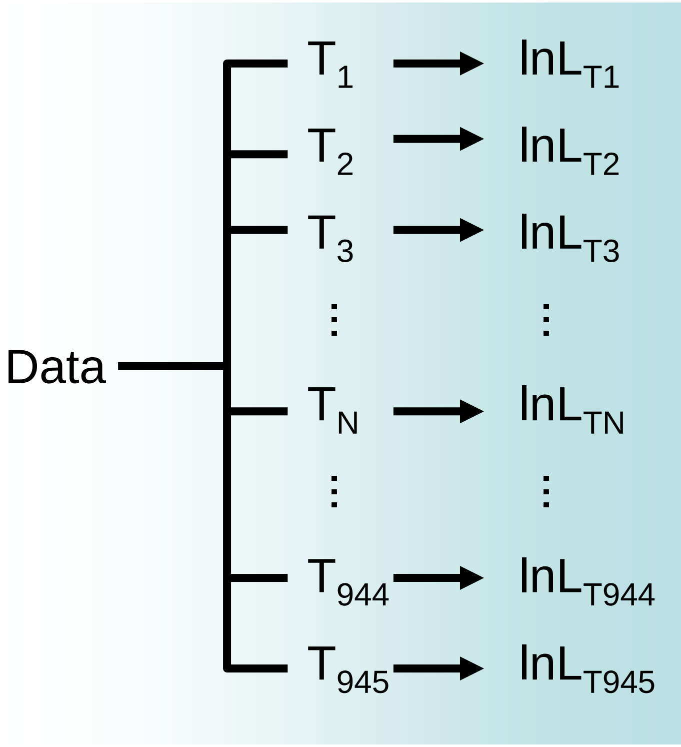
Analyses on PACS-CS

- To test the efficiency of HTS, we need to know where is the global peak
- Find the ML tree by exhaustive tree search
- Compare the ML tree and the trees from HTS
  - NJ→nni, NJ→spr, SWA→nni, & SWA→spr
  - Can HTS find the ML tree most efficiently?
- Reduce data size for ETS
  - 10 taxa
  - 2,027,025 trees to score

MPI-Puzzle

- 並列化：Drs. 中島、佐藤

Sequential



Data

T_1 → lnL_{T1}

T_2 → lnL_{T2}

T_3 → lnL_{T3}

⋮ ⋮

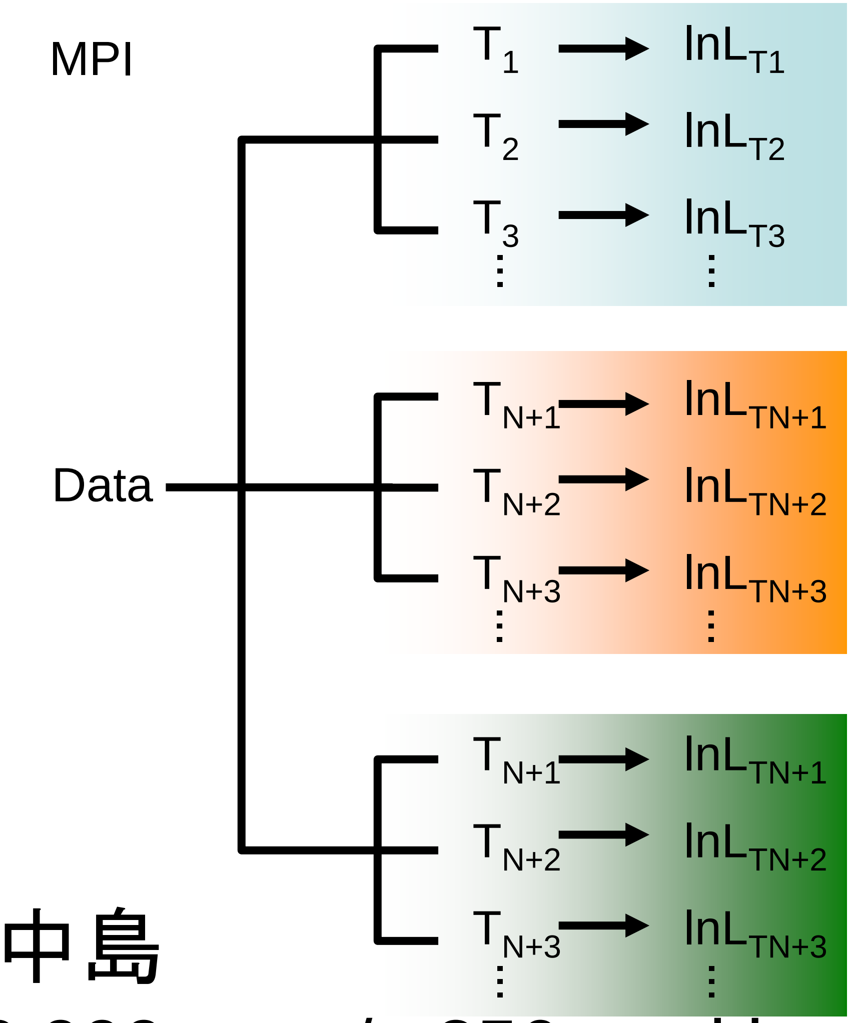
T_N → lnL_{TN}

⋮ ⋮

T_{944} → lnL_{T944}

T_{945} → lnL_{T945}

MPI



Data

T_1 → lnL_{T1}

T_2 → lnL_{T2}

T_3 → lnL_{T3}

⋮ ⋮

T_{N+1} → lnL_{TN+1}

T_{N+2} → lnL_{TN+2}

T_{N+3} → lnL_{TN+3}

⋮ ⋮

T_{N+1} → lnL_{TN+1}

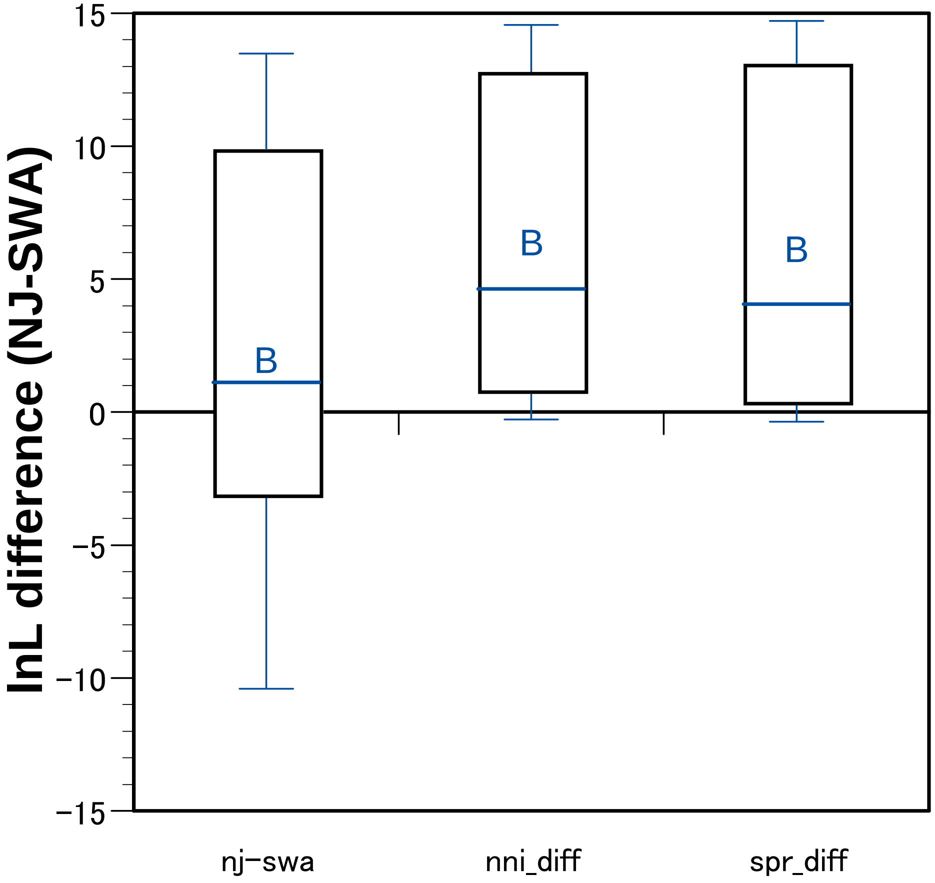
T_{N+2} → lnL_{TN+2}

T_{N+3} → lnL_{TN+3}

⋮ ⋮

- 実行: Drs. Choy & 中島
  - 0.5 hrs to score 100,000 trees/a 256 partition

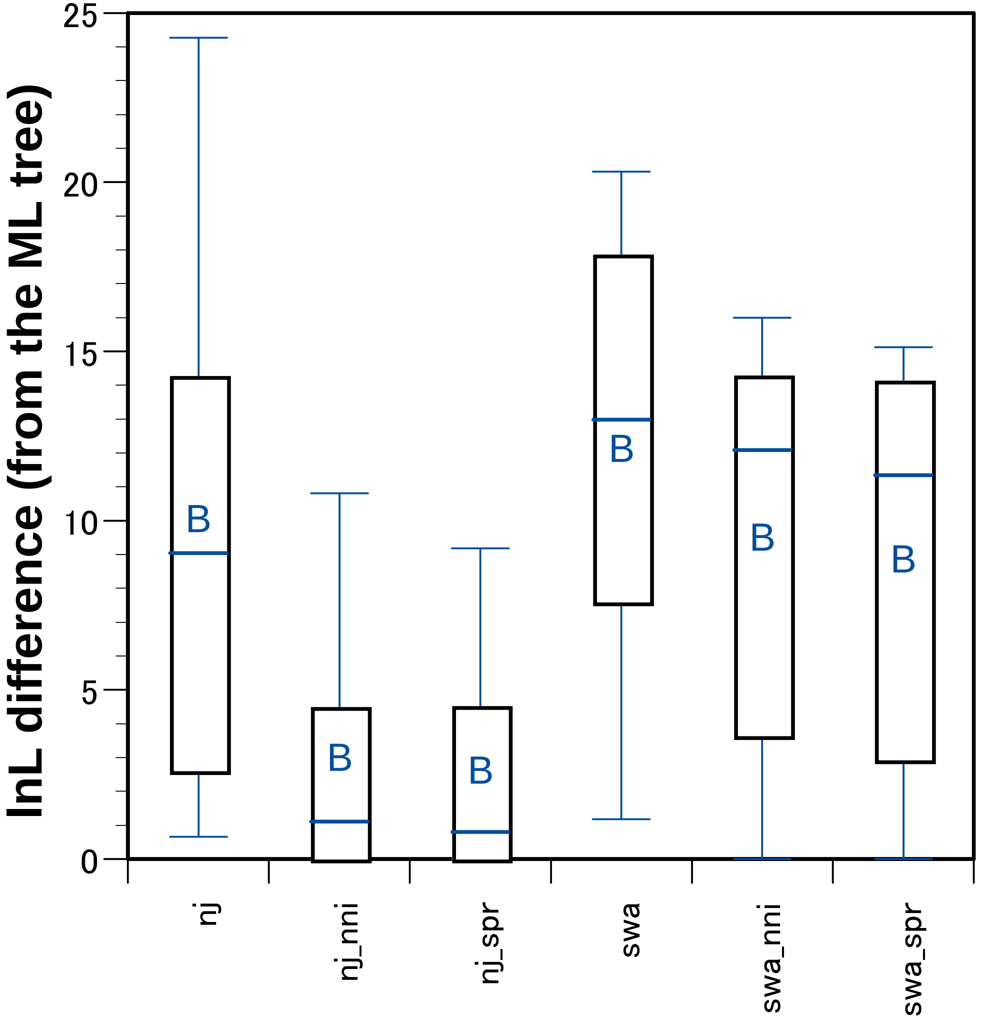
Start tree for HTS; NJ or SWA?



- "Best score" trees
  - NJ-based HTS
  - SWA-based HTS
- NJ-based HTS >> SWA-based HTS

- Did the NJ-based HTS migrate toward the ML trees?

Efficiency of HTS



- ML tree
  - – Selected from 2M trees by ETS
- Compared the ML tree and the “best score” trees from HTS
- How many times do “BS” trees hit to the ML trees?

“BS” trees are *bullshits*

Tree search	ML?	M+A	M-A	BS hit M+A…
ETS	n.a.	13	7	
HTS (nj→nni)	7			
HTS (nj→spr)	9			
HTS (swa→nni)	3			
HTS (swa→spr)	3			

- None of HTS was efficient
  - NJ-based HTS showed the best performance, but…

Conclusions (*so far*)

- lnL calculation of 2M trees has done for 20 data sets
- HTS cannot search the "*full* tree space"
  - Failed to find the ML trees in many cases
- The ML tree with M-A is difficult to find
  - NJ-based HTS gave an over-credit to M+A
  - SWA-based HTS tend to disfavor M+A

- Continue the 10-taxon ETS analyses
  - Repeat the HTS evaluation
  - Substitute Micro to other (short branch sequences)