An Overview: Parallel Phylogenetic Applications

Bui Quang Minh

Center for Integrative Bioinformatics Vienna

Max F. Perutz Laboratories

Computer Science Meeting

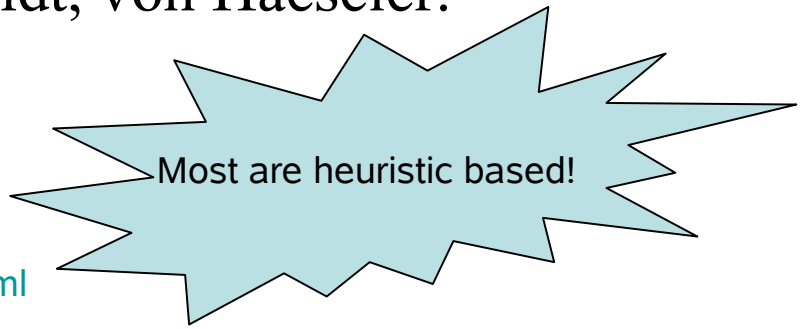
Vienna, 26th April 2006

Phylogenetic Methods

- Distance matrix
- Maximum parsimony
- Maximum likelihood
- Bayesian inference

Phylogenetic Applications

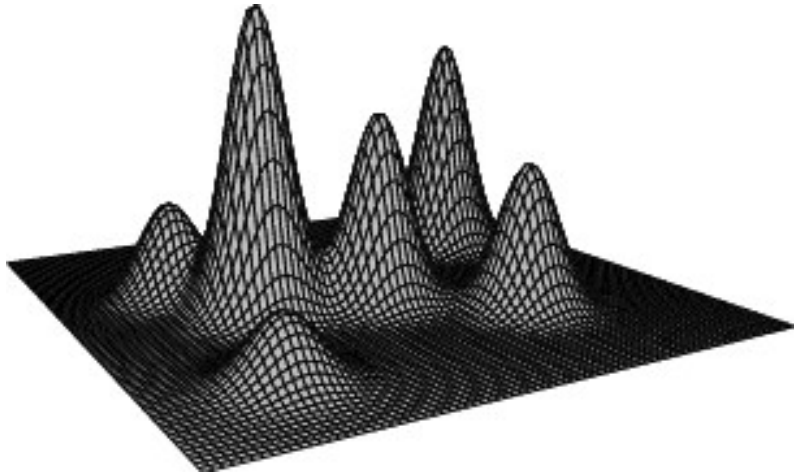
- fastDNAml (1994): Olsen et al.
- Tree-Puzzle (1996): Schmidt, Strimmer, Vingron, von Haeseler.
- MrBayes (2001): Altekar, Huelsenbeck, Ronquist.
- RAxML (2002): Stamatakis et al.
- PHYML (2003): Guindon & Gascuel.
- IQPNNI (2004): Minh, Vinh, Schmidt, von Haeseler.
- PHYLIP, PAUP*
- NJ, BIONJ,...



http://evolution.genetics.washington.edu/phylip/software.html
for a complete list.

Heuristics

- Stepwise Addition
- Local Rearrangement

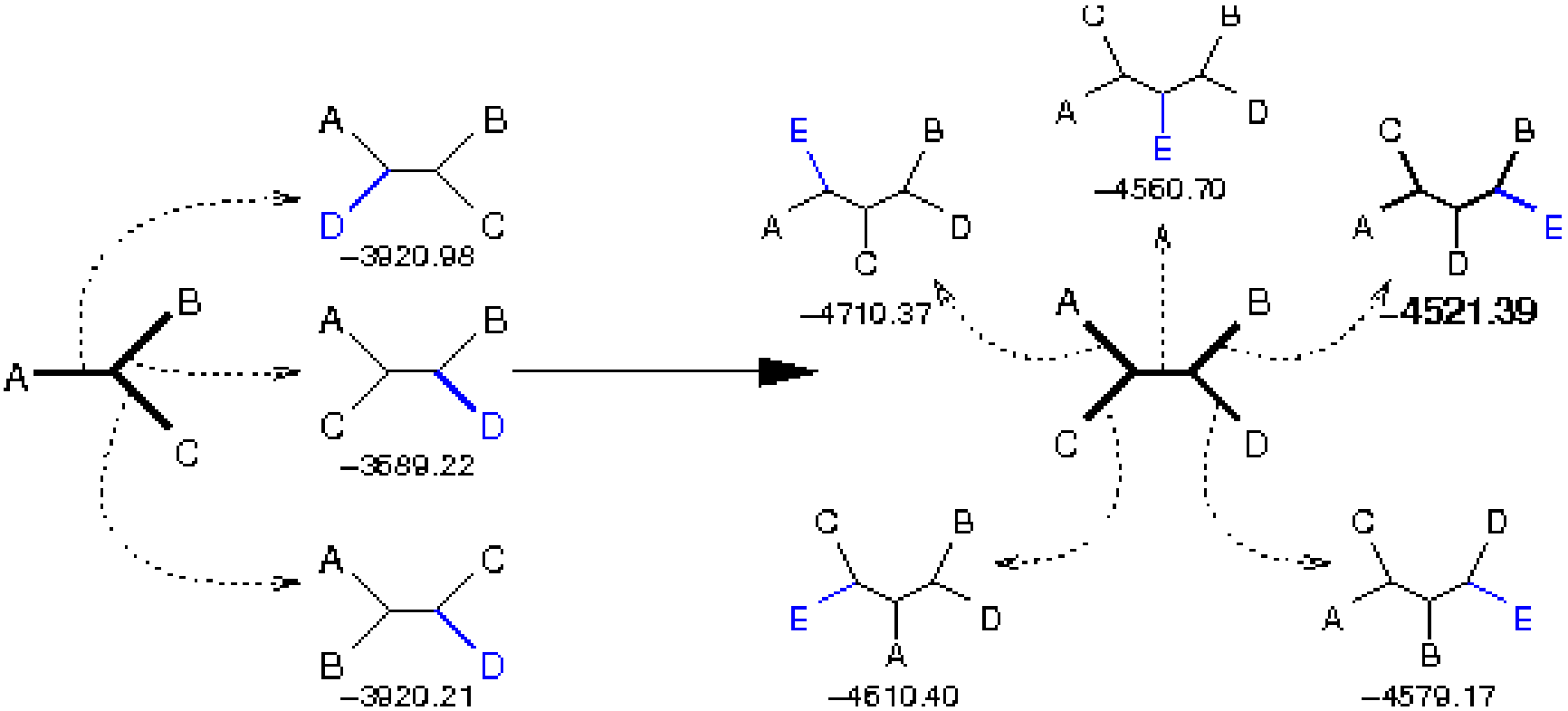


Greedy Algorithm!

Fig. Source: Heiko Schmidt

Stepwise Addition

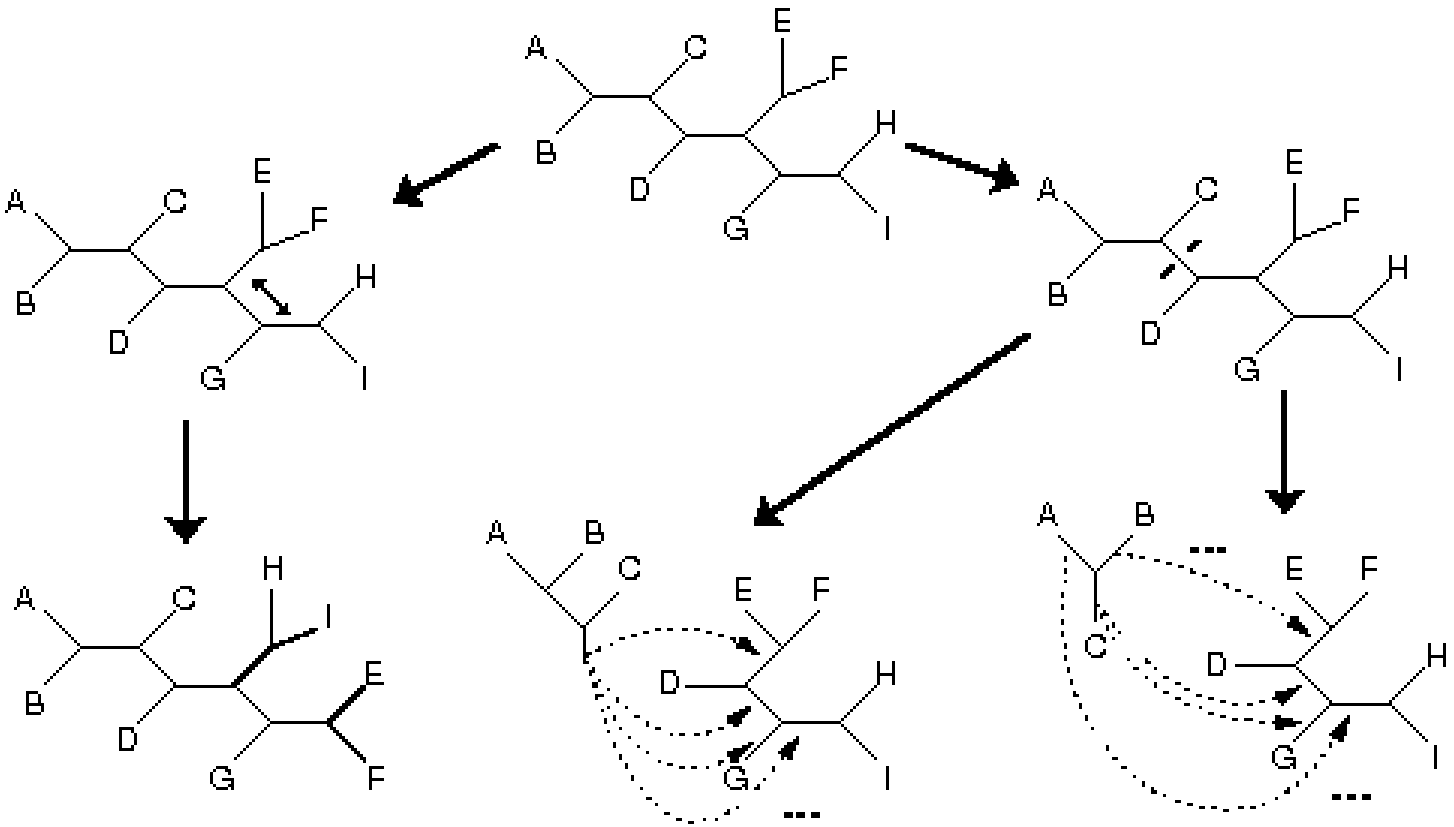
Used in fastDNAml, Tree-puzzle, RAxML, PHYML, IQPNNI



Randomized order of inserted taxa

Fig. Source: Heiko Schmidt

Local Rearragement



Nearest Neighbor Interchange

Possible NNI trees = O(n)

Used in PHYML,IQPNNI

subtree pruning + regrafting

Possible SPR trees = O(n*n)

Used in fastDNAml, RAxML

tree–bisection + reconnection

Possible TBR trees = O(n*n*n)

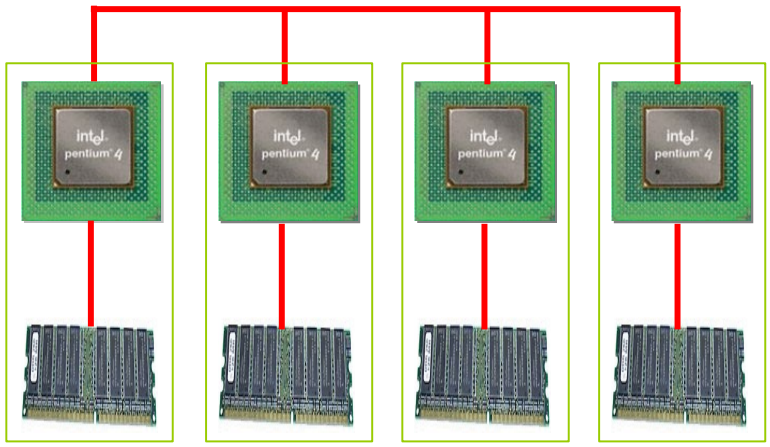
Used in PAUP*

Fig. Source: Heiko Schmidt

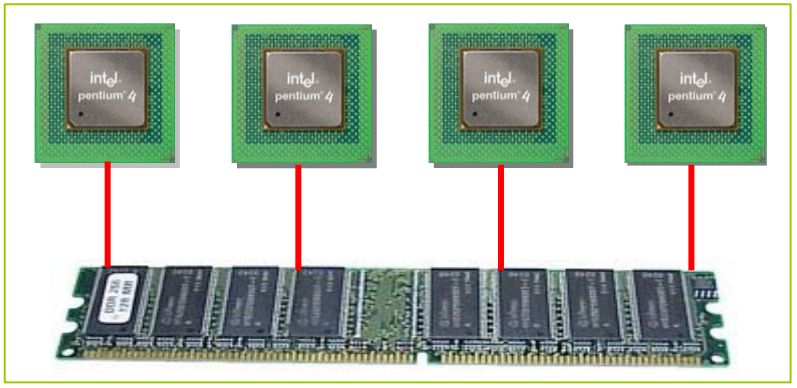
„The most promising consideration, however, is parallel computing.“

Trelles (2001) On the parallelisation of bioinformatics applications.

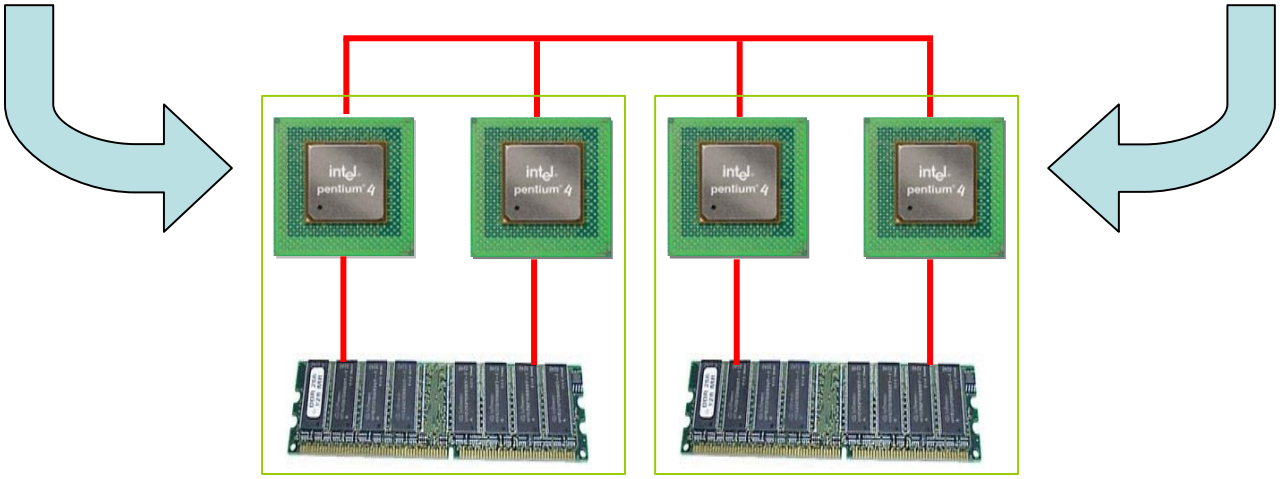
Parallel Computing



Distributed-memory: MPI, PVM



Shared-memory (SMP): OpenMP, Cashmere

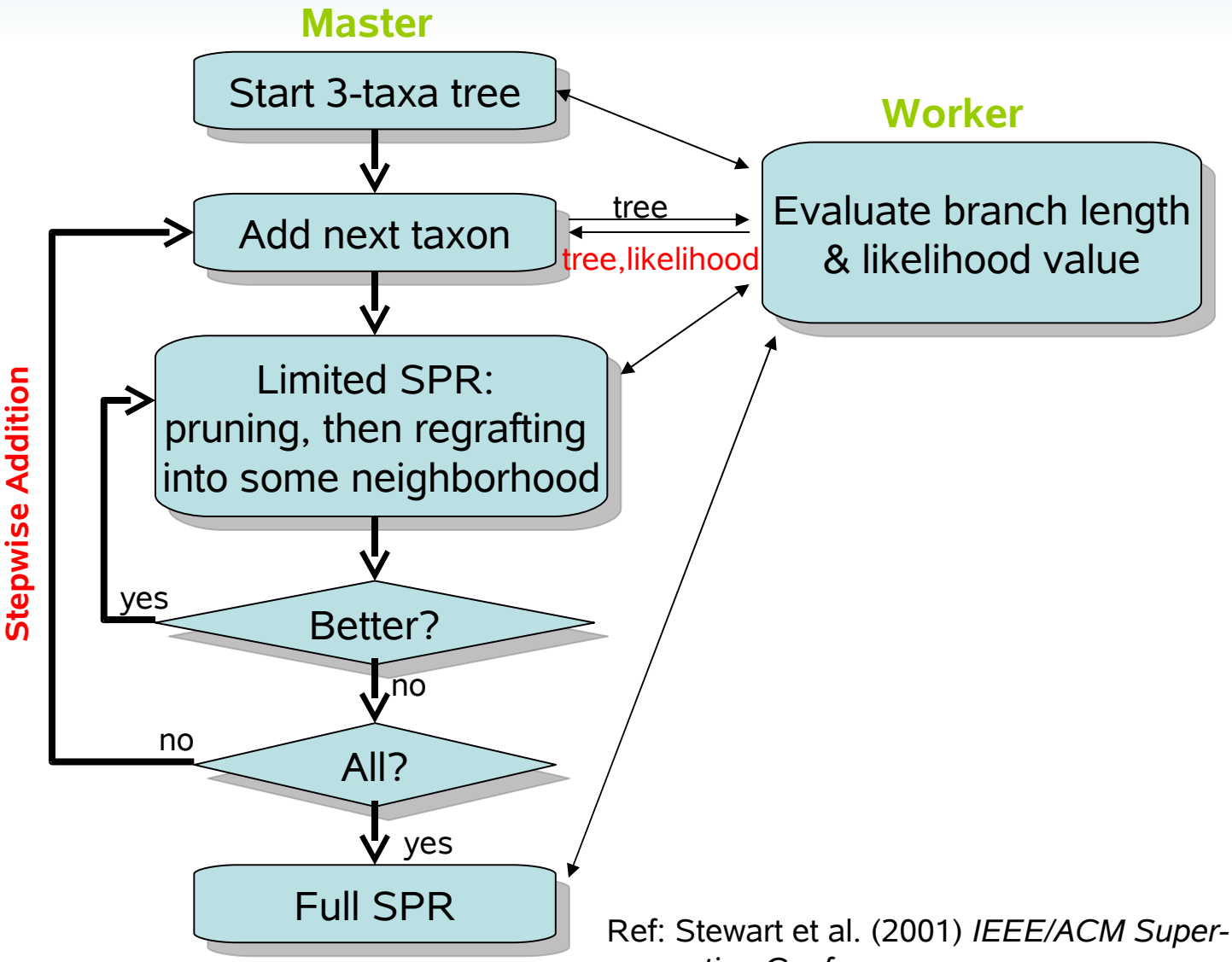


Cluster of SMPs: hybrid MPI/OpenMP

Available parallel programs

- fastDNAml (1994): P4 (1994), MPI, PVM (2001)
- Tree-Puzzle (1996): MPI (2002)
- MrBayes (2001): MPI, Cashmere (2004)
- RAxML (2002): MPI, OpenMP (2003,2005)
- IQPNNI (2004): MPI, OpenMP, hybrid MPI/OpenMP (2005)

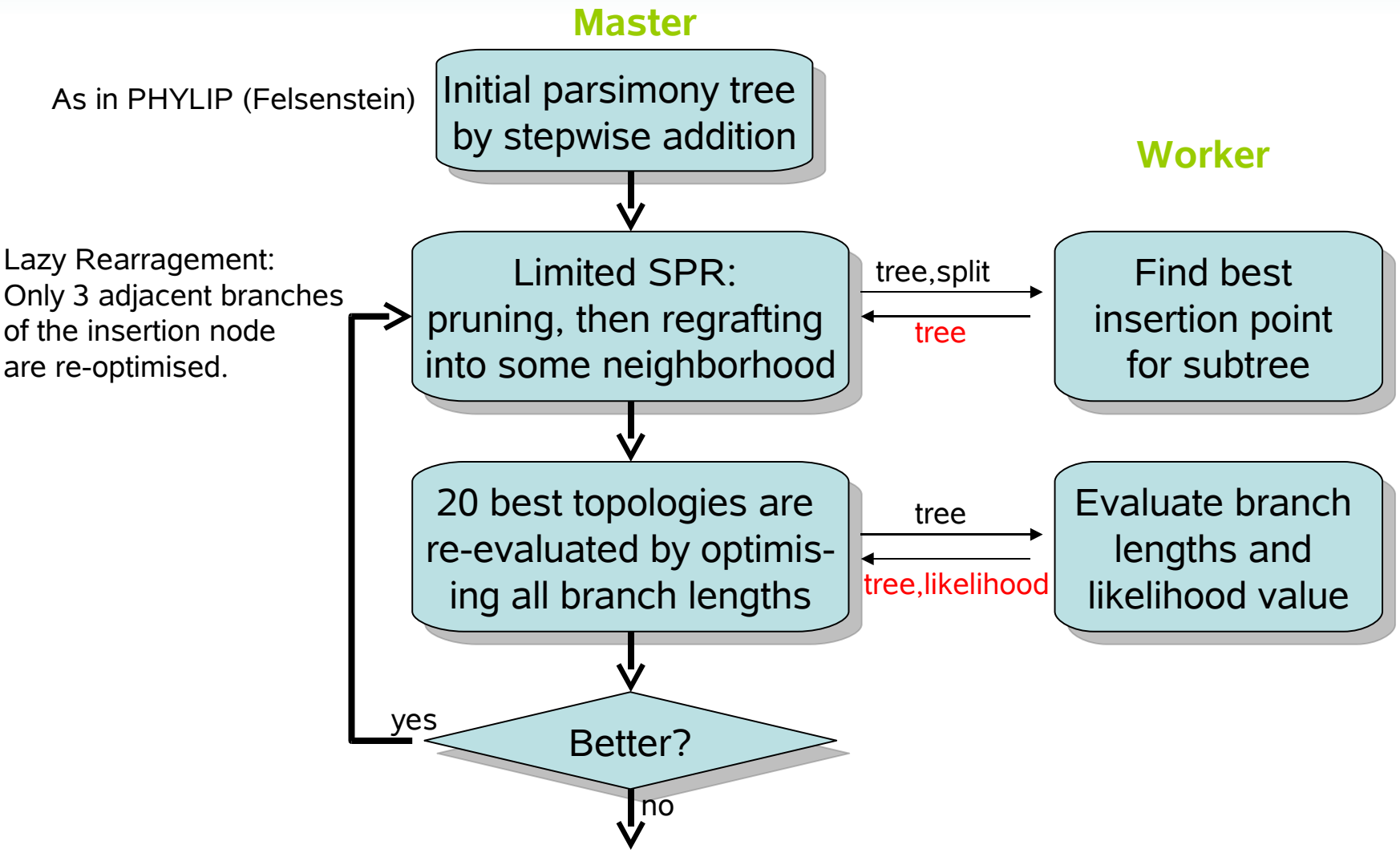
fastDNAml



Random Order of added taxon!

Ref: Stewart et al. (2001) *IEEE/ACM Super-computing Conference*.

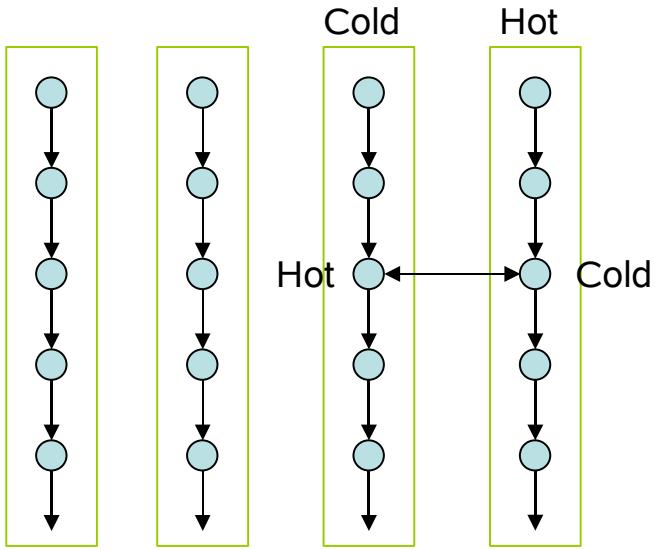
RAxML



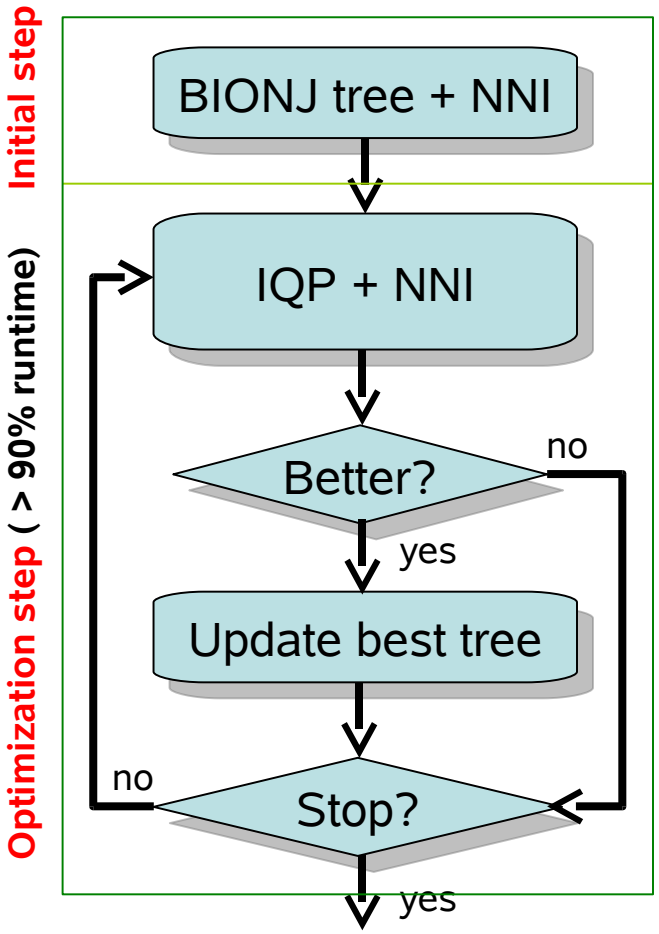
Ref: Stamatakis et al. (2003-2005)

MrBayes

- Maximum Likelihood:
$$L(Tree) = \Pr(Data|Tree)$$
- Bayesian Inference:
$$\Pr(Tree|Data) = \Pr(Data|Tree)\Pr(Tree)/\Pr(Data)$$
- Metropolis-Coupled Markov Chain Monte Carlo (MC^3)



IQPNNI



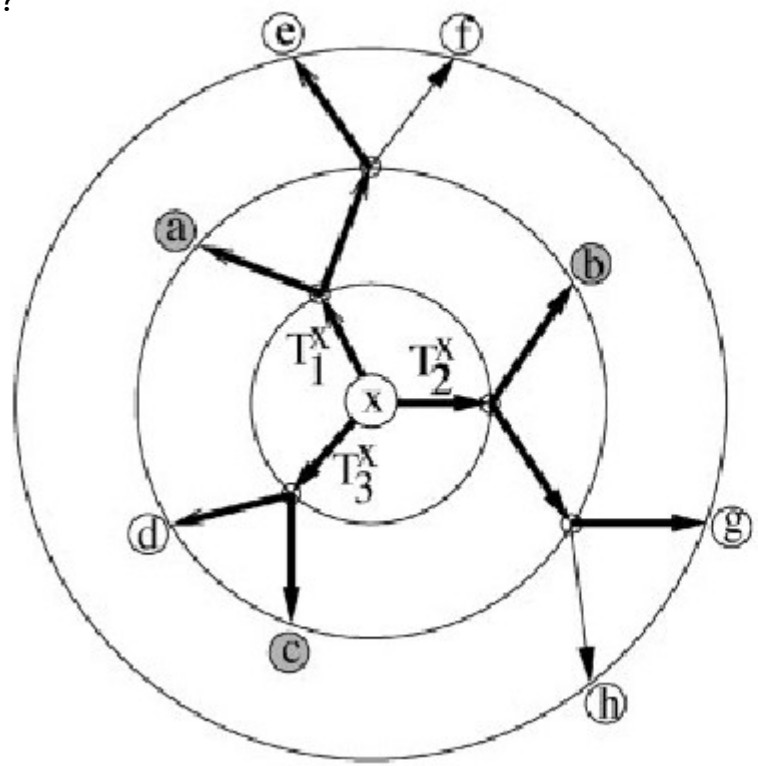
- Combination of heuristics
  - **BIONJ**: Neighbor Joining.
  - **IQP**: Important Quartet Puzzling.
  - **NNI**: Nearest Neighbor Interchange.
- Stop condition
  - A number of iterations.

Ref: Vinh LS, von Haeseler A (2004) IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, 21,1565-71.

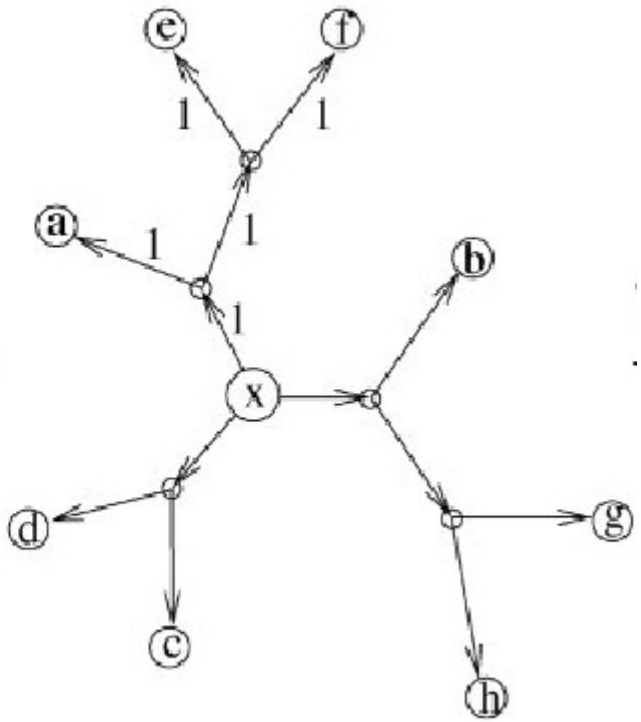
IQP: Important Quartet Puzzling

1. Randomly remove p_{del} leaves from tree.
2. Stepwise add deleted leaves into tree by:

(y)?

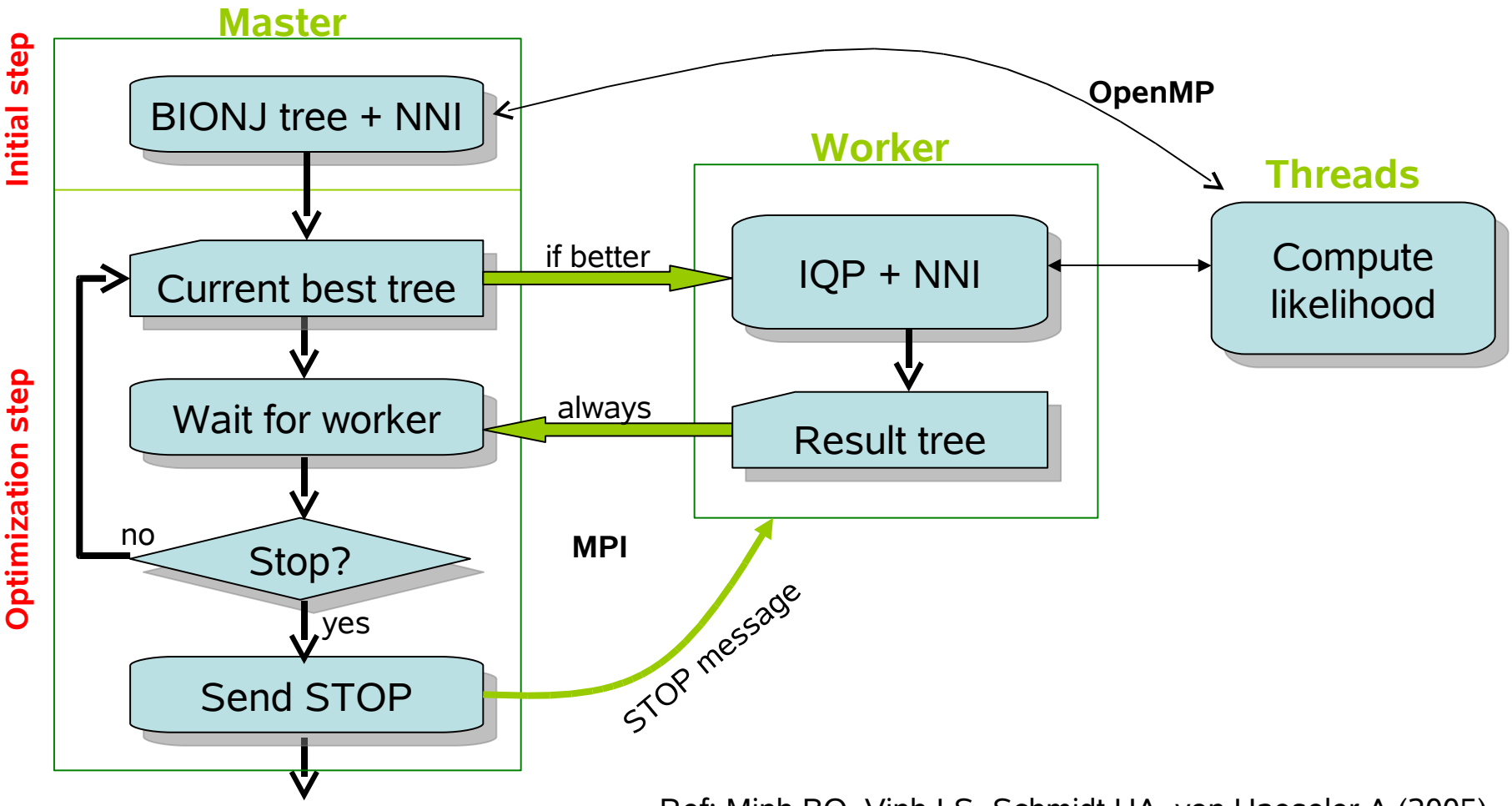


$T_{ay \mid bc}$



$S_2(T_1^x)=\{a,e\}$, $S_2(T_2^x)=\{b,g\}$, $S_2(T_3^x)=\{c,d\}$

Parallel IQPNNI



Ref: Minh BQ, Vinh LS, Schmidt HA, von Haeseler A (2005) pIQPNNI - Parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, 21(19), 3794-6.

IQPNNI: OpenMP

Tree likelihood calculation

```
for (site=0; site < nsites; site++) {
      logli_site = compute_log_likelihood(site);
      logli += logli_site;
}
```

These for-loops consume > 90% of runtime!

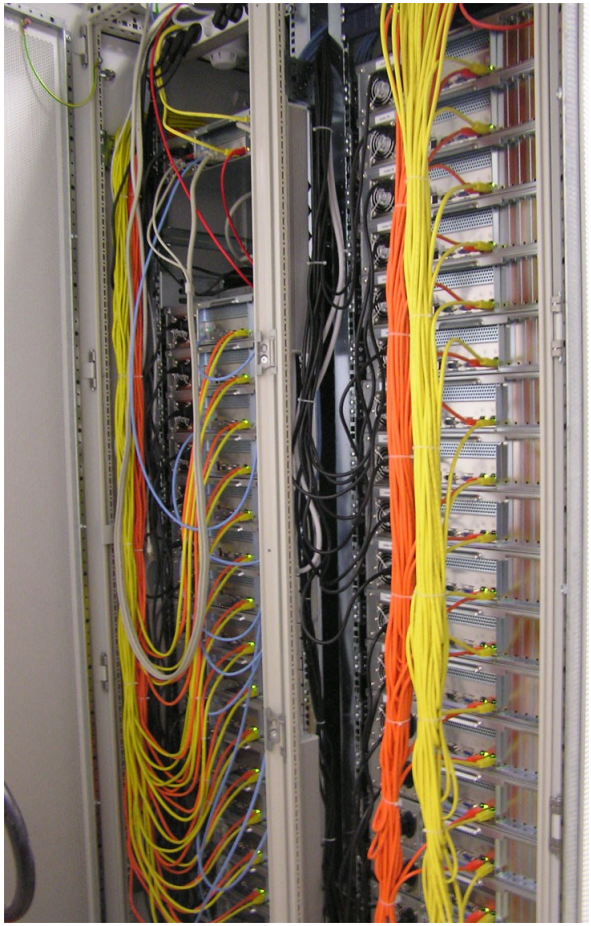
```
#pragma omp parallel for private(logli_site) \
      reduction(+: logli)
for (site=0; site < nsites; site++) {
      logli_site = compute_log_likelihood(site);
      logli += logli_site;
}
```

Datasets

Dataset	Name	Model	#Sequences	#Sites	#Iterations
1000dna	sim.	HKY85	1000	1000	300
500dna	rbcL[1]	-	500	1398	300
218dna	ssu rRNA[2]	-	218	4182	300
74aa	mito74[3]	WAG	74	4013	200

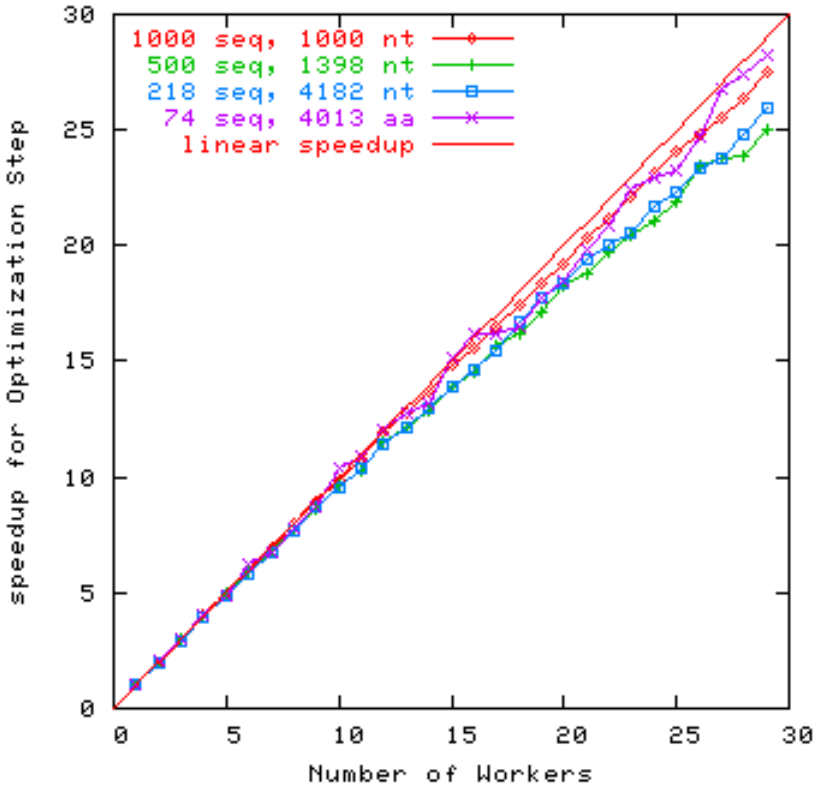
[1] from plant plastids.
[2] prokaryotic sequences from the small ribosomal subunit.
[3] unpublished alignment of verterbrate sequences from Schmidt HA, Liebers D.

Düsseldorf´s system



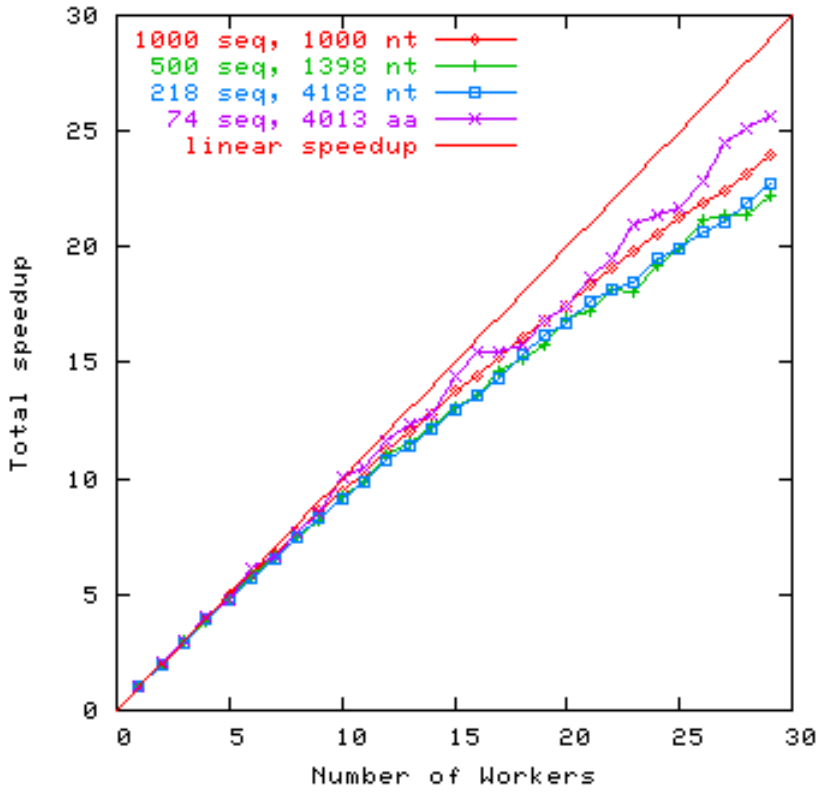
- 20x2 AMD Opteron 2.0 GHz.
- 18x2 AMD Opteron 1.8 GHz.
- **Test speedup on 30 CPUs of 2.0 GHz.**

Pure MPI speedup

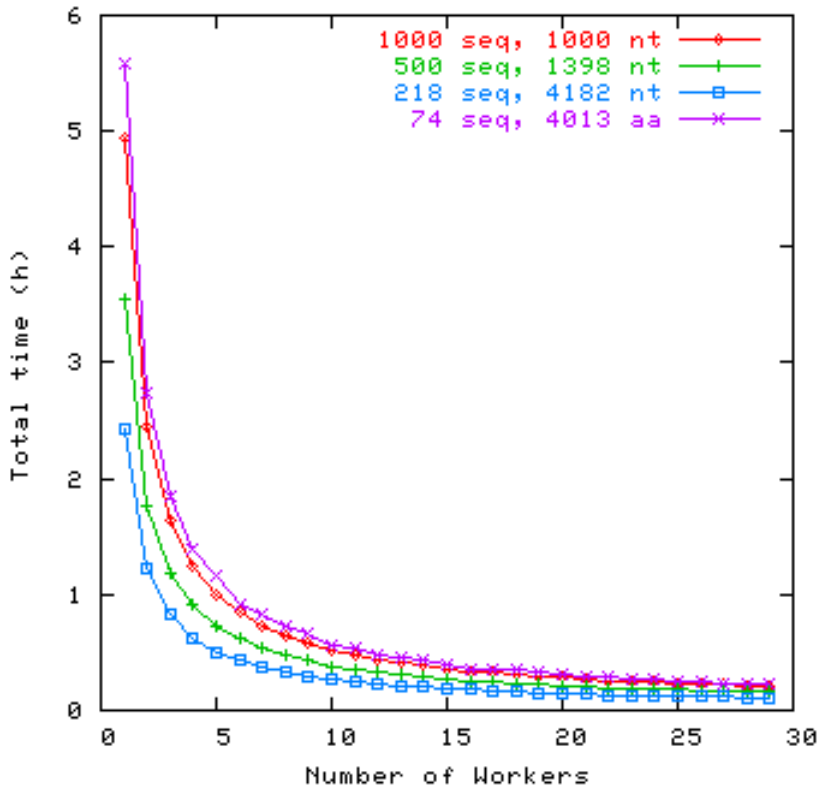


Optimization step

Pure MPI total speedup



Total speedup



Total running time

Jülich´s system

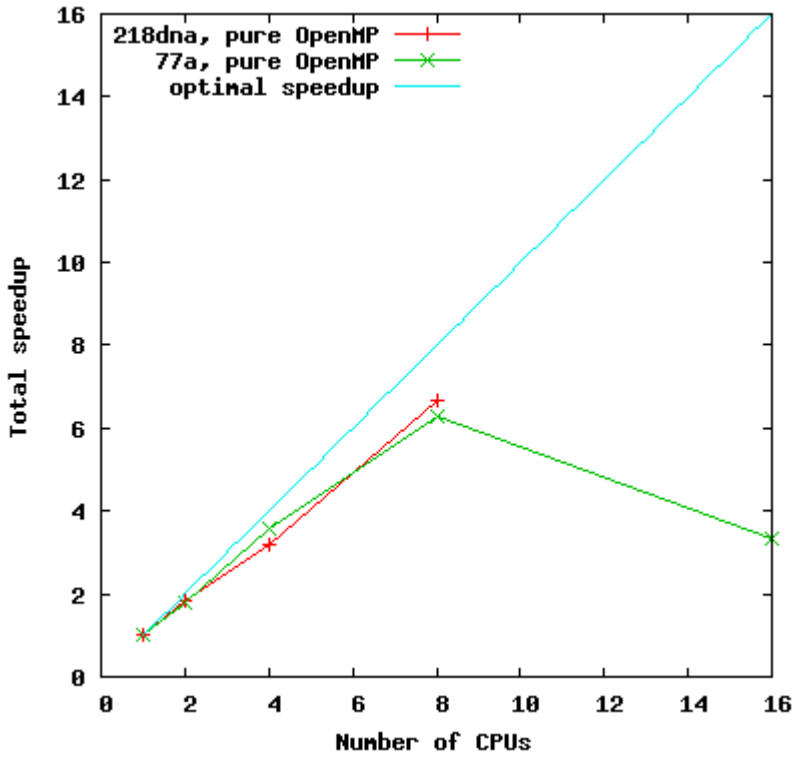
- Jülich Multi Processor (JUMP): IBM eServer pSeries 690.
- 41 SMPs, 32 CPUs each.
- Power4+ 1.7 GHz.
- **Test up to 128 CPUs.**

Datasets

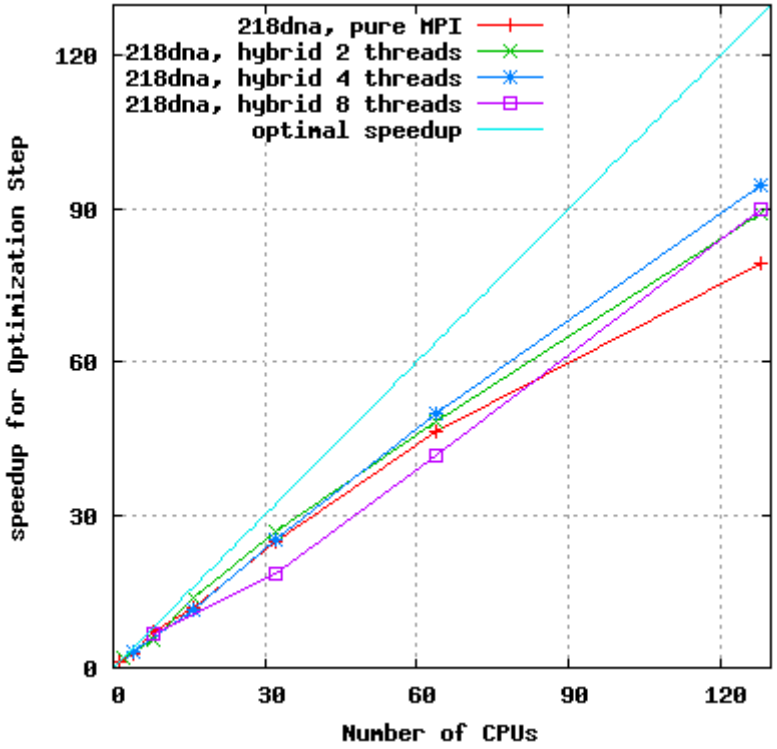
Dataset	Name	Model	#Sequences	#Sites	#Iterations
1000dna	sim.	HKY85	1000	1000	300
500dna	rbcL[1]	-	500	1398	300
218dna	ssu rRNA[2]	-	218	4182	**150**
74aa	mito74[3]	WAG	74	4013	**50**

[1] from plant plastids.
[2] prokaryotic sequences from the small ribosomal subunit.
[3] unpublished alignment of verterbrate sequences from Schmidt HA, Liebers D.

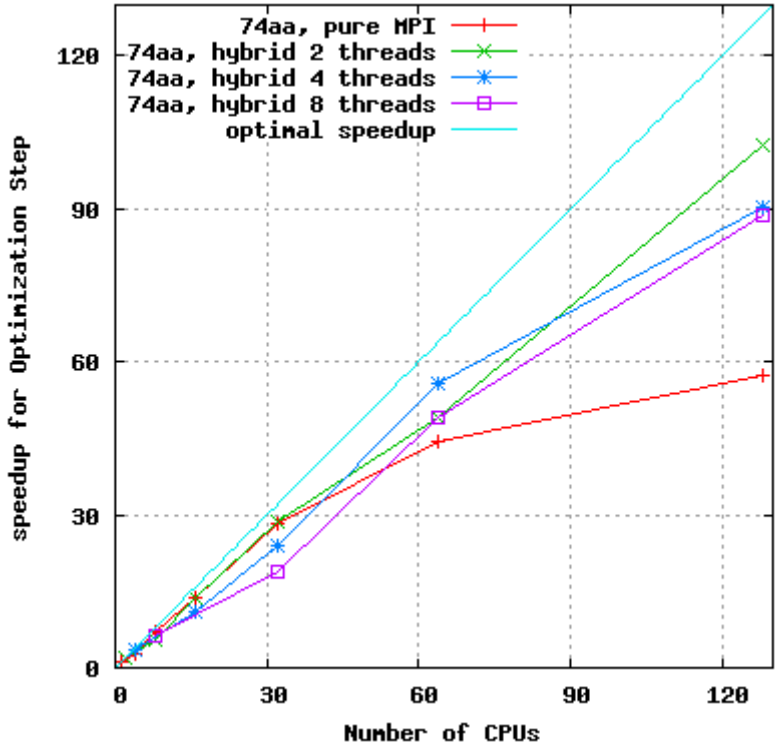
Pure OpenMP speedup



Pure MPI vs. Hybrid speedup: Optimization Step

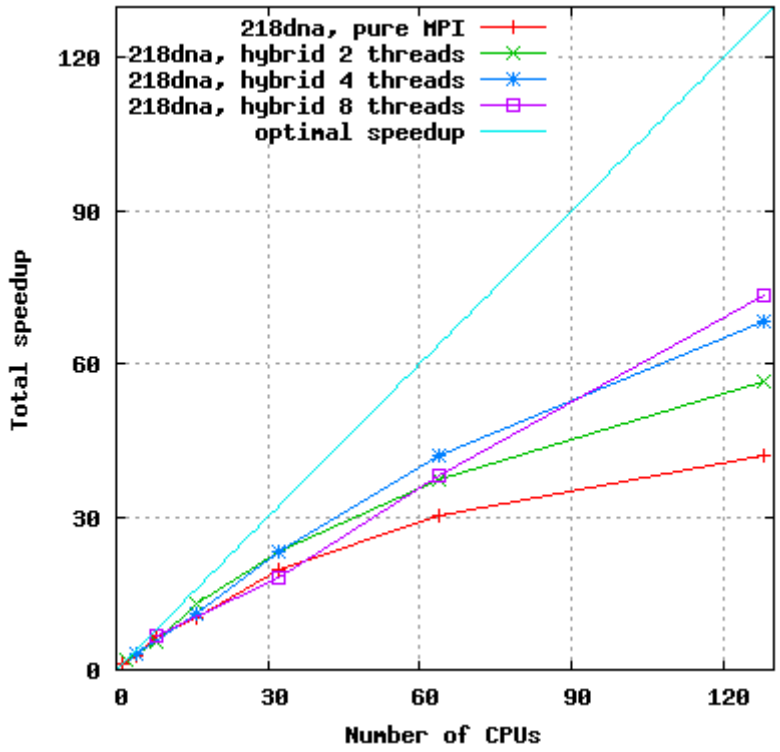


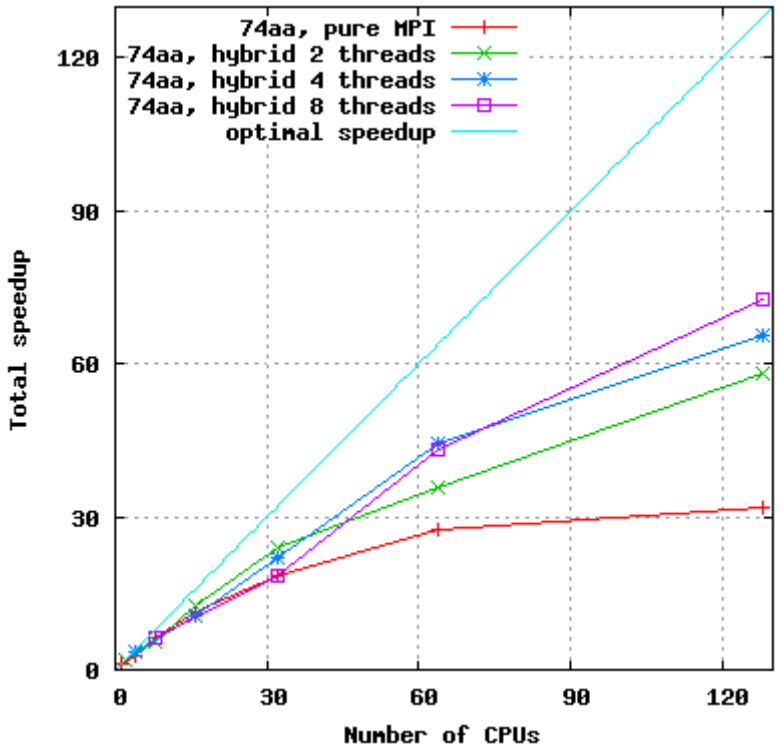
Dataset 218dna



Dataset 74aa

Pure MPI vs. Hybrid total speedup





Runtime 218dna on 128 CPUs

#Processes	#Threads	Inital step	Opt. step	Total
128	pure MPI	34s	36s	70s
64	2	20s	32s	52s
32	4	13s	30s	43s
16	8	08s	32s	40s

Sequential runtime 49m:05s

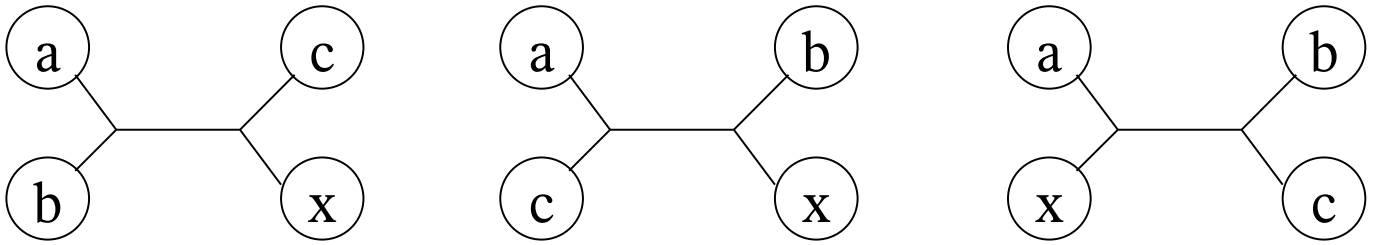
Discussion

- Other optimizations: genetic algorithm (GA), simulated annealing (SA), ant colony optimization (ACO).
- A Standard Phylogenetic Library?
- Parallel, Distributed Platform.

Thanks to...

- Arndt von Haeseler
- Vinh Le Sy
- Heiko A. Schmidt
- Marc-Andre Hermanns
- Other colleagues
- Computing facilities at NIC/ZAM (Jülich)
- Your attention!

Quartet Puzzling



- **Quartet tree: tree of four leaves.**
- **Stepwise addition algorithm:**
  - **Initialize with any three-leaf tree.**
  - **Insert leaf *x* into current tree:**
    - **Take any 3 leaves a, b, c from current tree.**
    - **Find optimal topology among $T_{ab|cx}$, $T_{ac|bx}$, $T_{ax|bc}$.**
    - **Insert x into branch which is supported by most quartets.**

Likelihood framework

- Likelihood(tree) = Probability(data|tree).
- Find a tree which maximizes its likelihood.
- Each site evolves independently.
  - $P(\text{data}|\text{tree}) = P(\text{site}_1|\text{tree}) \times \ldots \times P(\text{site}_m|\text{tree})$.
  - $\log\text{-likelihood}(\text{tree}) = \log P(\text{site}_1|\text{tree}) + \ldots + \log P(\text{site}_m|\text{tree})$.