RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees

A. Stamatakis[1,], T. Ludwig[2] and H. Meier[1]*

[1]Department of Computer Science, Technical University of Munich, Boltzmannstrasse 3, D-85748 München, Germany and [2]Department of Computer Science, University of Heidelberg, Im Neuenheimer Feld 348, D-69120 Heidelberg, Germany

Presented by

Pei-Chen Peng

April 23, 2015

Introduction

- The number of tree topologies grows exponentially with the number of taxa.
- Maximum likelihood recovers the true tree more frequently than neighbor joining and parsimony.
- ML methods were limited to 100 taxa as of 2005.

Related work

- **MrBayes** outperforms all other analyzed phylogeny programs in terms of speed and tree topology (Williams and Moret, 2003).
  - Simulated data set with 60 taxa.
- **PHYML** outperforms other recent approaches including MrBayes (Guindon and Gascuel, 2003).
  - Simulated data set with 100 taxa.
  - Real data sets with 218 and 500 taxa.

Goal

- Develop a rapid ML-based phylogeny inference method for large evolutionary trees.
- Compare with MrBayes and PHYML.

RAxML-III---Initilization

- RAxML-III starts by a parsimony tree.
  - Obtains better likelihood value compared to random or neighbor joining trees.
  - Builds a tree rapidly. RAxML-III can executes several times with different starting trees.

RAxML-III---Optimization (1)

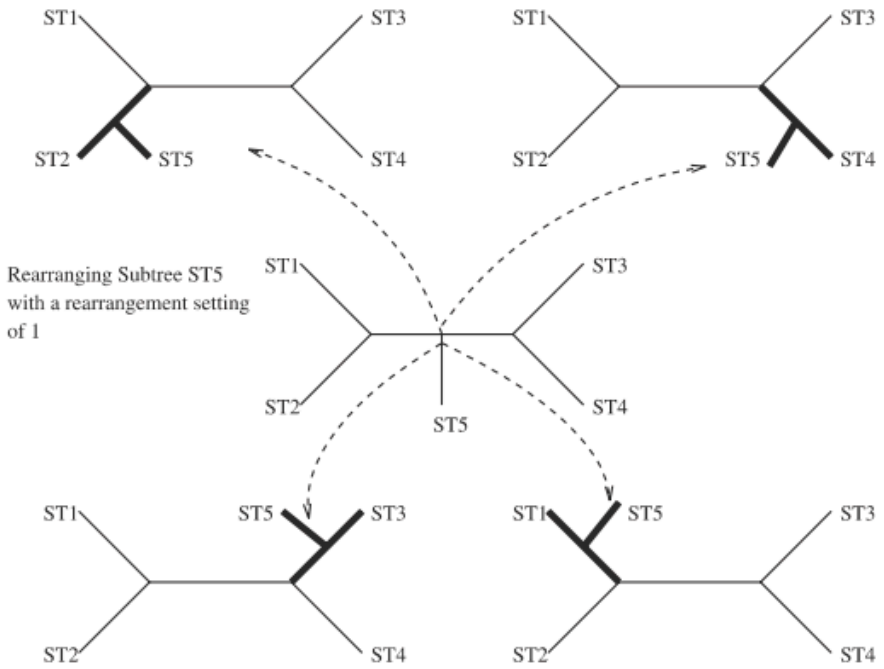

Fig. 1. Rearrangements traversing one node for subtree ST5, branches which are optimized by RAxML-III are indicated by bold lines.

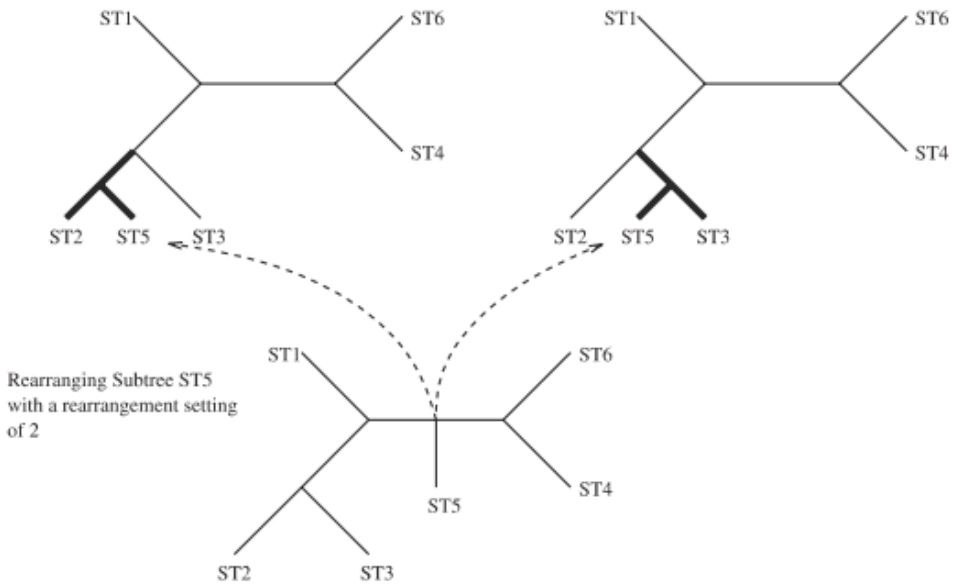

Fig. 2. Example rearrangements traversing two nodes for subtree ST5, branches which are optimized by RAxML-III are indicated by bold lines.

RAxML-III---Optimization (2)

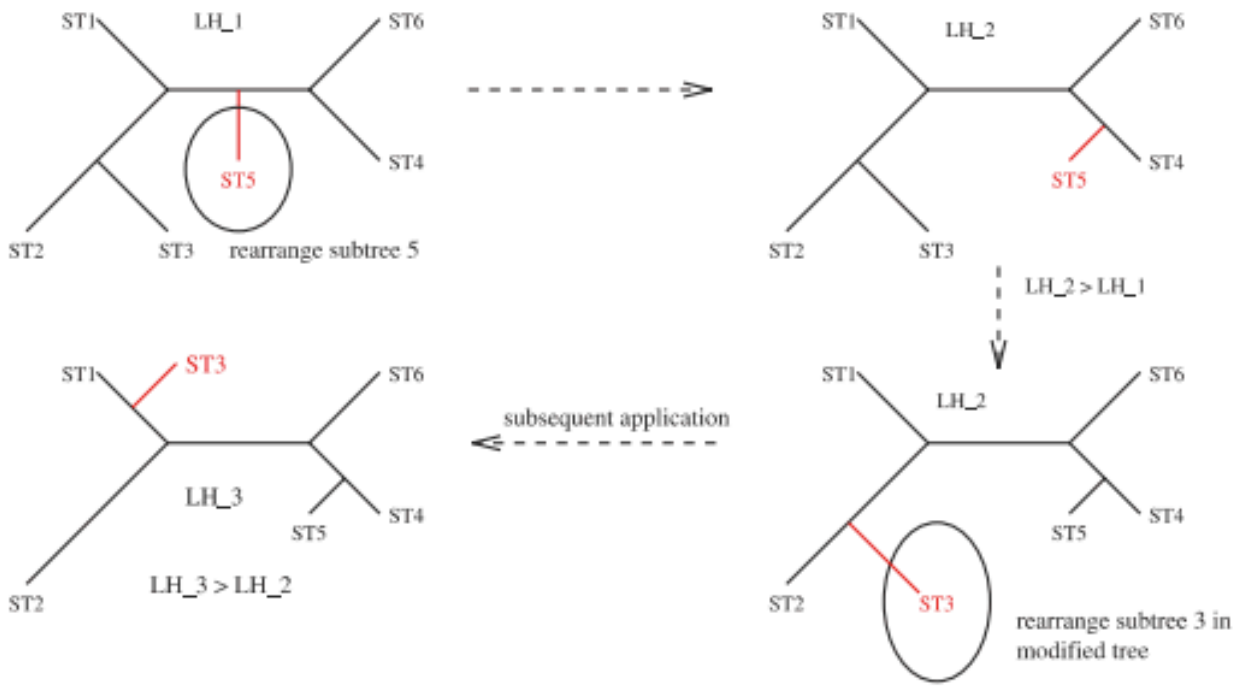

Fig. 3. Example for subsequent application of topological improvements during one rearrangement step.

RAxML-III--- Termination

- *rL* : minimum distance of moving subtrees

  rU : maximum distance of moving subtrees
- As long as the tree doesn't improve, increase *rL* and *rU* by *rStart*.
- RAxML-III terminates when *rU* >= *rMax*.
- Build a consensus tree from the set of final trees.

Performance on simulated data sets

Table 2. Topological accuracy and execution times for PHYML & RAxML-III on simulated data

data	PHYML (RF)	secs	RAxML (RF)	secs
4000_SIM_1	0.065	18944	0.065	9152
4000_SIM_2	0.039	22273	0.037	50609
4000_SIM_3	0.033	24907	0.027	97962
4000_SIM_4	0.030	30870	0.031	85080
4000_SIM_5	0.028	24182	0.035	91178
4000_SIM_6	0.027	32614	0.031	176686
4000_SIM_7	n/a	n/a	0.028	144519
4000_SIM_8	0.027	34750	0.032	185454
4000_SIM_9	0.026	18828	0.036	78061
4000_SIM_10	n/a	n/a	0.034	64690

Performance on real data sets

Table 3. PHYML, MrBayes, RAxML-III execution times and likelihood values for real data sets

data	PHYML	secs	MrBayes	secs	RAxML	secs	R > PHY	secs	PAxML	hrs
101_SC	−74097.6	153	−77191.5	40527	−73919.3	617	−74046.9	31	−73975.9	47
150_SC	−44298.1	158	−52028.4	49427	−44142.6	390	−44262.9	33	−44146.9	164
150_ARB	−77219.7	313	−77196.7	29383	−77189.7	178	−77197.6	67	−77189.8	300
200_ARB	−104826.5	477	−104856.4	156419	−104742.6	272	−104809.0	99	−104743.3	775
250_ARB	−131560.3	787	−133238.3	158418	−131468.0	1067	−131549.4	249	−131469.0	1947
500_ARB	−253354.2	2235	−263217.8	366496	−252499.4	26124	−252986.4	493	−252588.1	7372
1000_ARB	−402215.0	16594	−459392.4	509148	−400925.3	50729	−401571.9	1893	−402282.1	9898
218_RDPII	−157923.1	403	−158911.6	138453	−157526.0	6774	−157807.9	244	n/a	n/a
500_ZILLA	−22186.8	2400	−22259.0	96557	−21033.9	29916	−22036.9	67	n/a	n/a

Performance on real data sets (2)

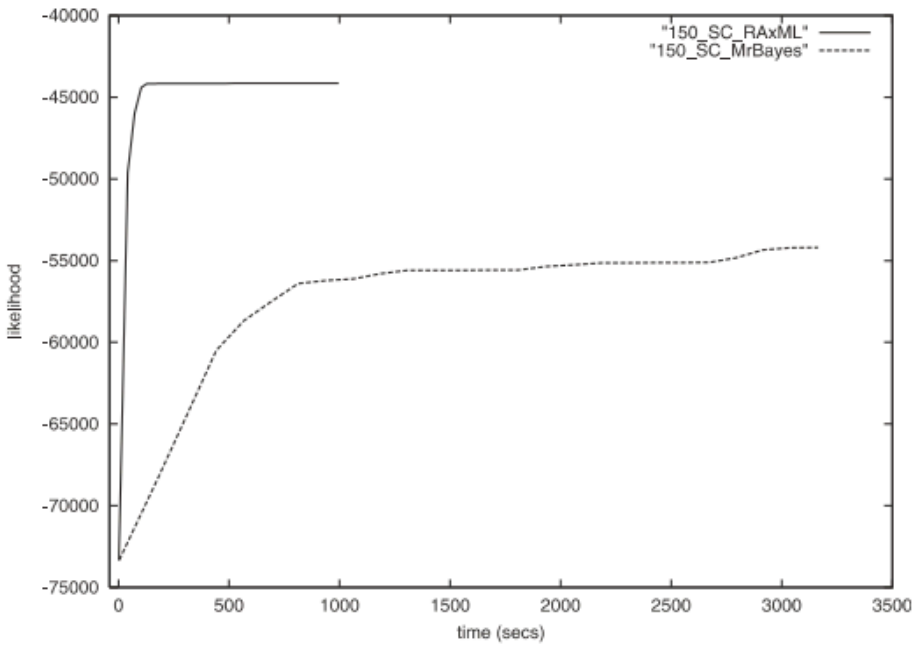

Fig. 6. 150_SC likelihood improvement over time of RAxML-III and MrBayes for the same random starting tree.

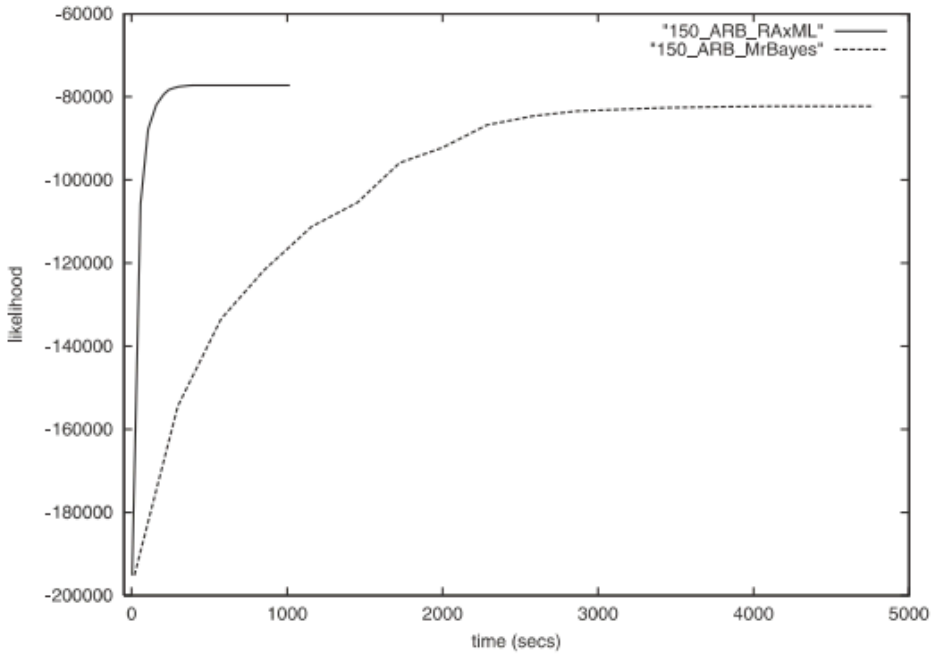

Fig. 7. 150_ARB likelihood improvement over time of RAxML-III and MrBayes for the same random starting tree.

Performance on real data sets (2)

Table 4. Performance of PHYML and RAxML-III for HKY85 and GTR models of evolution with model parameter optimization

data	PHYML	secs	RAxML	secs	R > PHY
HKY85					
101_SC	−74035	104	−73908	71	21
150_SC	−44315	85	−44219	64	26
150_ARB	−76881	190	−76863	94	49
200_ARB	−104316	282	−104270	185	120
250_ARB	−131013	405	−130926	342	116
500_ARB	−252224	1453	−251781	1049	420
1000_ARB	−400881	3908	−399732	3633	1666
2025_ARB	−372746	9749	−371472	8426	4779
218_RDPII	−156895	230	−156663	331	126
GTR					
101_SC	−73814	131	−73638	119	49
150_SC	−44139	132	−44043	157	60
150_ARB	−76500	235	−76490	203	144
200_ARB	−103789	714	−103758	352	262
250_ARB	−130518	526	−130353	416	218
500_ARB	−250858	1170	−250238	1516	688
1000_ARB	−398731	4727	−397612	5731	2469
2025_ARB	−370539	5299	−369197	10771	5558
218_RDPII	−155881	316	−155748	406	268

Conclusion

- RAxML-III (on real data) and PHYML (on simulated data) represent very fast and accurate conventional maximum likelihood programs.

Thank You