Scaling up Tree Reconstruction methods to 100K Taxa and Beyond

Alexandros Stamatakis
Junior Research Group Leader

The Exelixis Lab
Bioinformatics Unit
Department of Computer Science
Technical University of Munich

stamatak@cs.tum.edu
http://wwwkramer.in.tum.de/exelixis

Dataset Shapes?

Badly shaped

10,000 bp
30,000 –
50,000 taxa

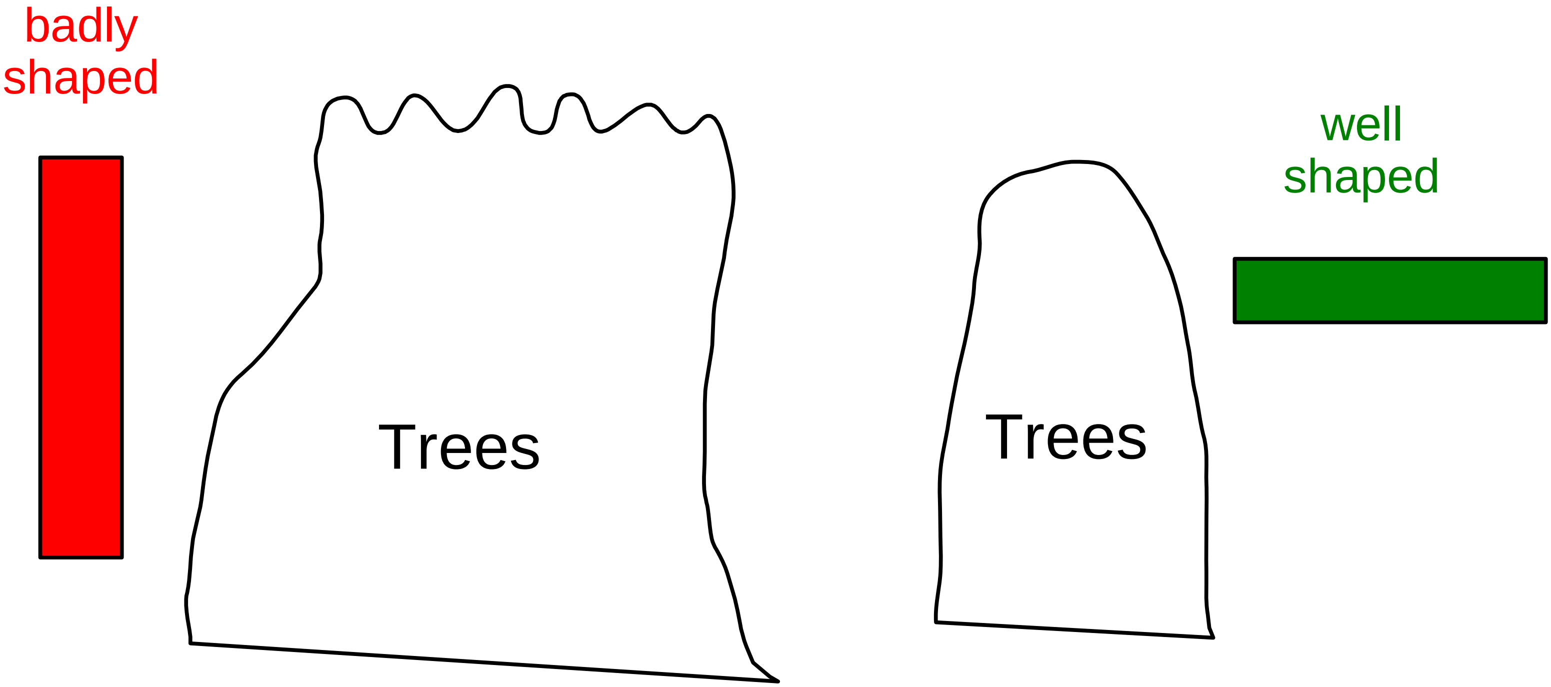
- Different parallelization strategies required
- Different search strategies required

Orangutan	A A C G T T T T -
Gorilla	A A G G T T T - -
Chimp	A - G G T T T T -
Homo Sapiens	A G G A T T T T T

Well shaped

1,000,000 bp
100 taxa
2.25 million CPU
Hours
110 GB of RAM

Easy & Difficult Likelihood Surfaces



Rough likelihood surface:
many taxa, few bp

Smooth likelihood surface:
Few taxa, many bp

Working Group

- WC Lead: Alexandros Stamatakis
- HPC expert: Wayne Pfeiffer at SDSC
- PhD Students (DFG): **Fernando Izquierdo**, Simon Berger, Nikos Alachiotis, Michael Ott
- Neighbor Joining & Approximate Likelihood (FastTree): Morgan Price LBNL
- Link to data assembly: Stephen Smith, Casey Dunn

Strategy

- Assuming badly-shaped DNA data
- Initially, simplify, re-write, & optimize parts of RAxML
- Separate into three distinct parts:
  1) Alignment parsing component
  2) Maximum Parsimony starting tree component
  3) ML component with CAT approximation of rate heterogeneity
- Optimize MP and ML kernels
- Improve algorithms
- De-Novo Parallelization

dependency

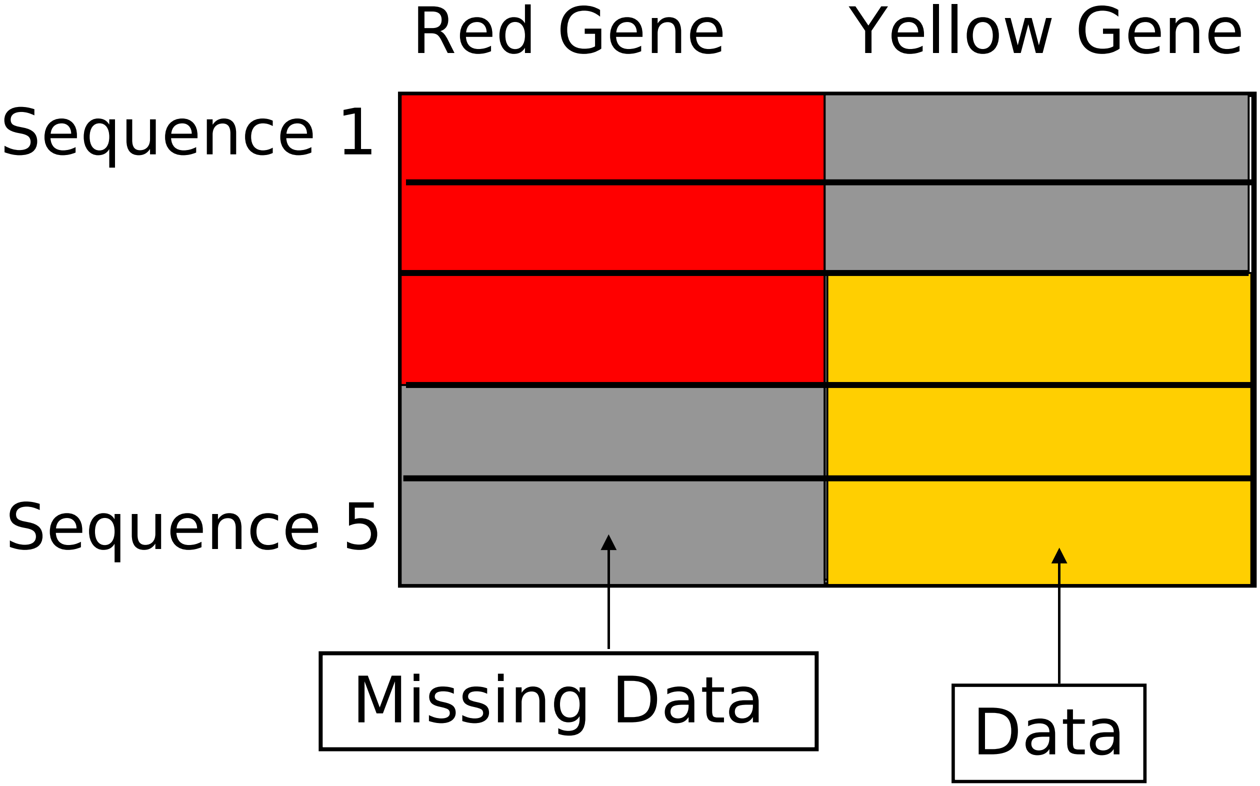
MP Kernel Optimization

- Completely re-designed MP kernel
- Bit-level operations
- SSE3 operations
- 2000 taxa single gene: stepwise addition + a couple of SPRs: 290.5 secs → 22.00 secs factor **13** !
- **Deliverable: before Christmas**

ML Kernel Optimization

- Adapt Likelihood function to gappy datasets
  - 36K and 55K we analyzed with Stephen are gappy
  - except for rbcL there isn't much data
- Reduce the number of FLOPs required to compute likelihood on a tree
- Reduce memory footprint proportional to sampling-induced gappyness

Phylogenomic Alignments



Red Gene

Yellow Gene

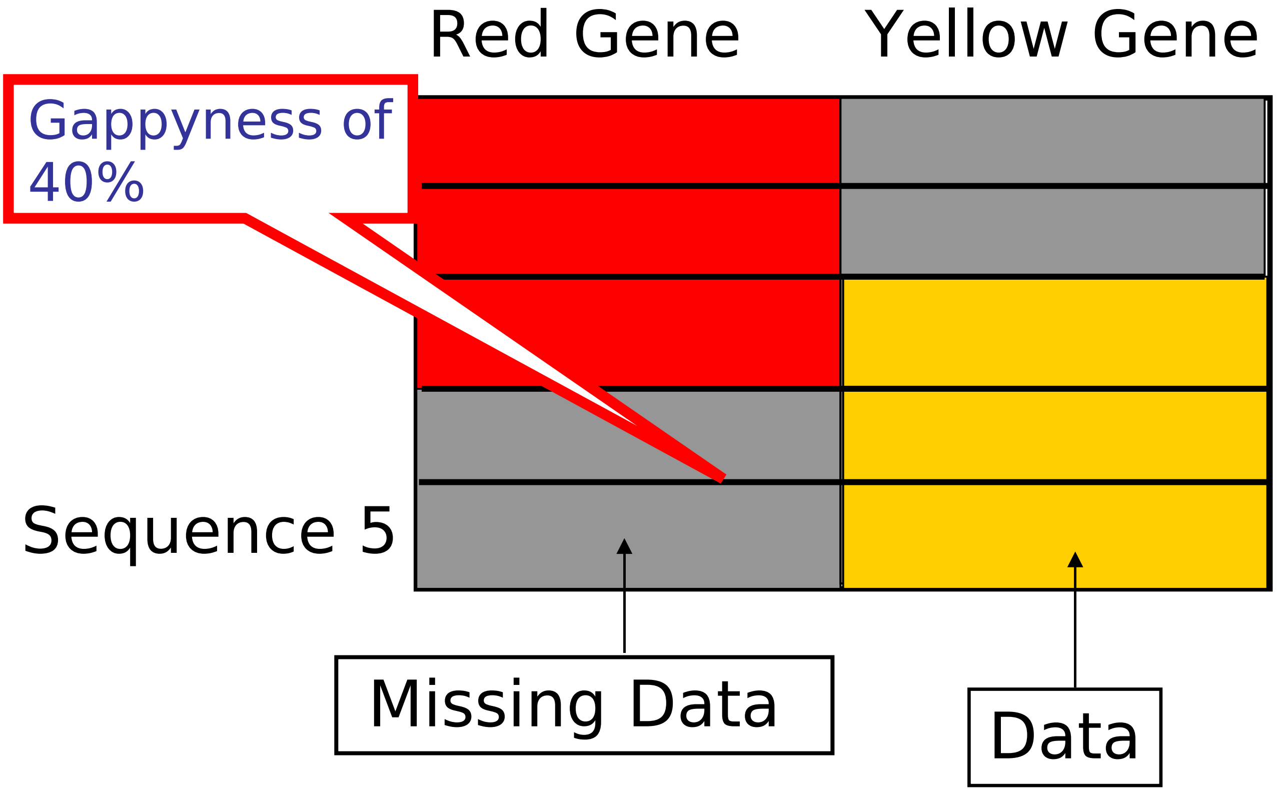
Sequence 1

Sequence 5

Missing Data

Data

Phylogenomic Alignments



Red Gene

Yellow Gene

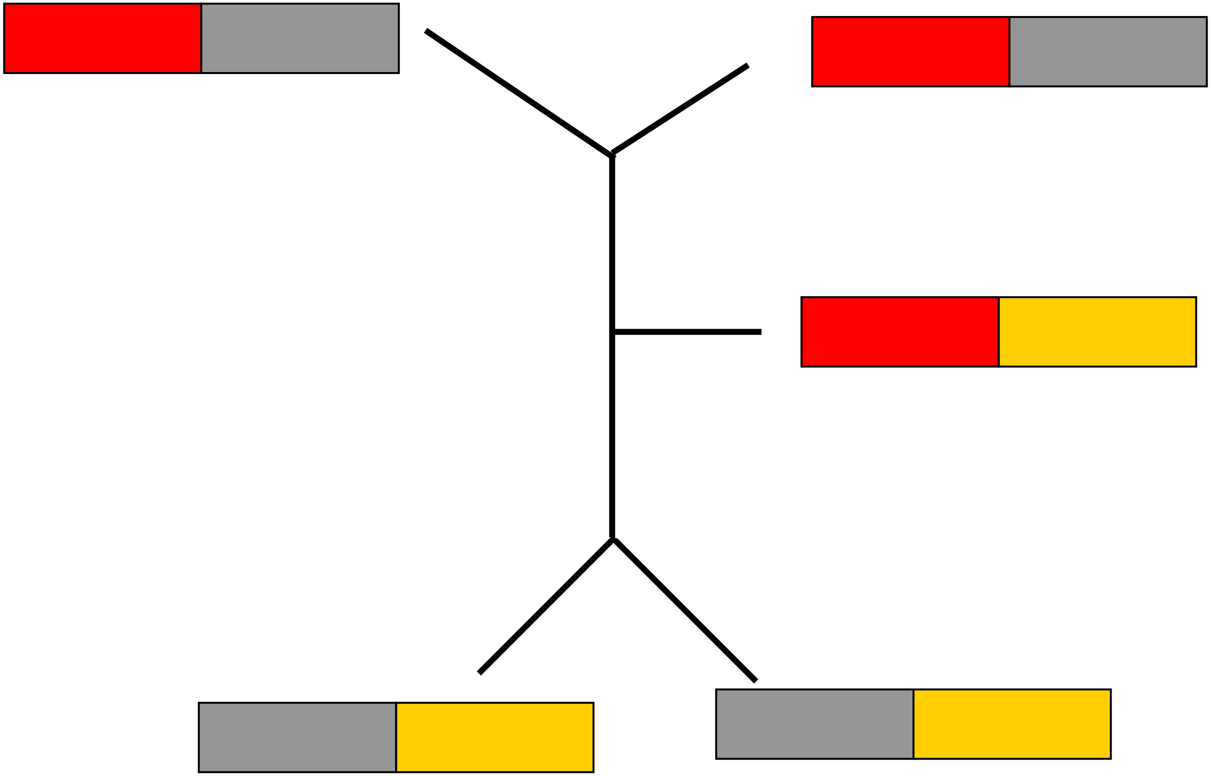
Gappyness of 40%

Sequence 5

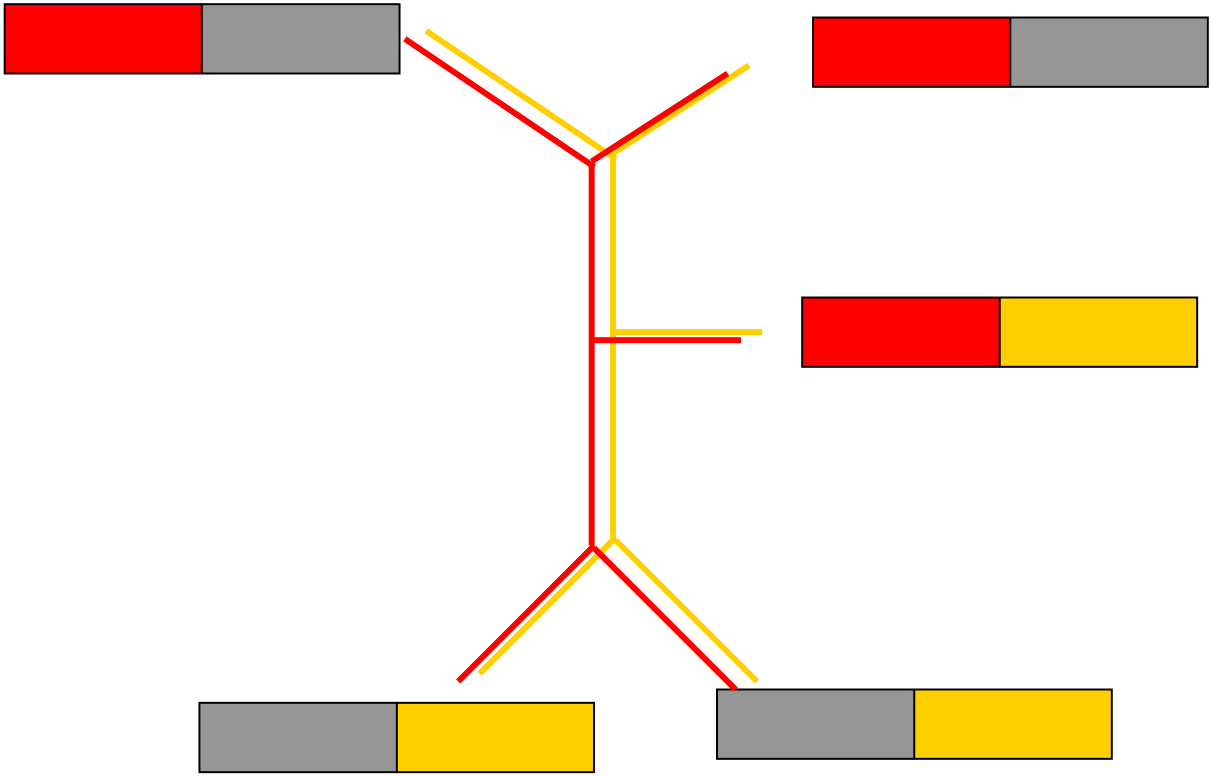
Missing Data

Data

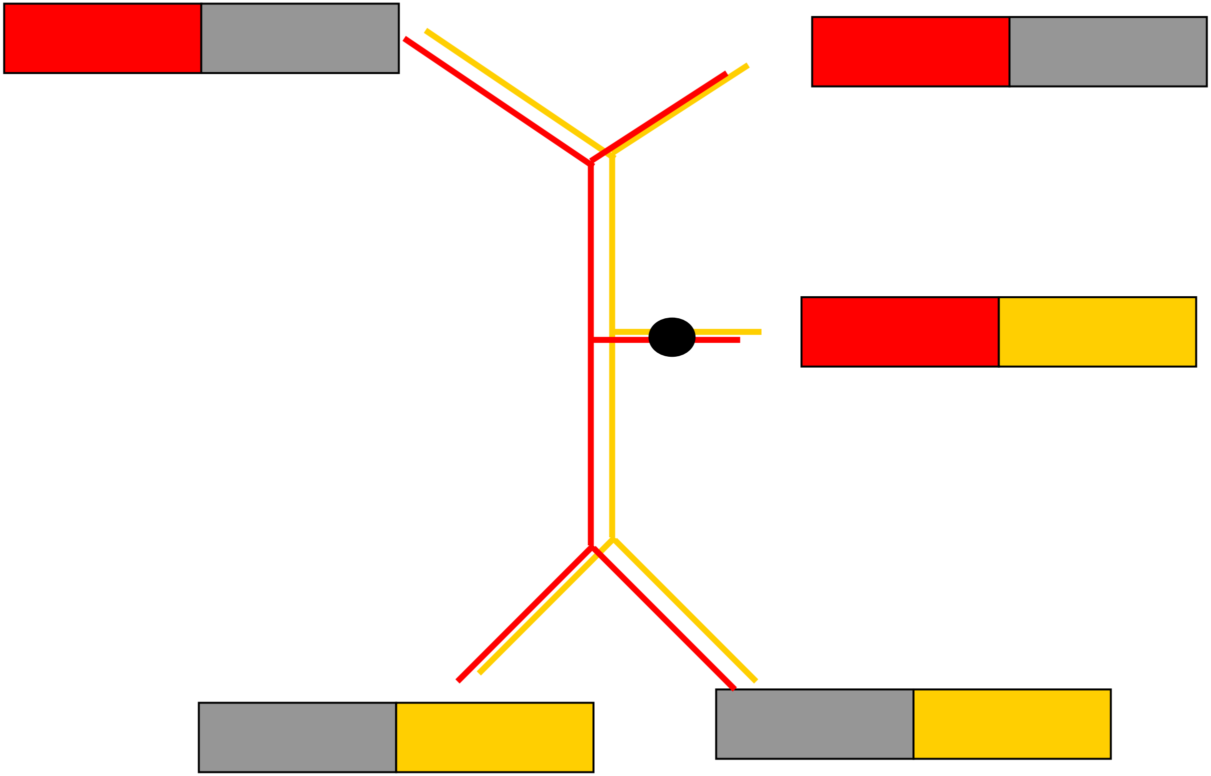
A Multi-Gene Model



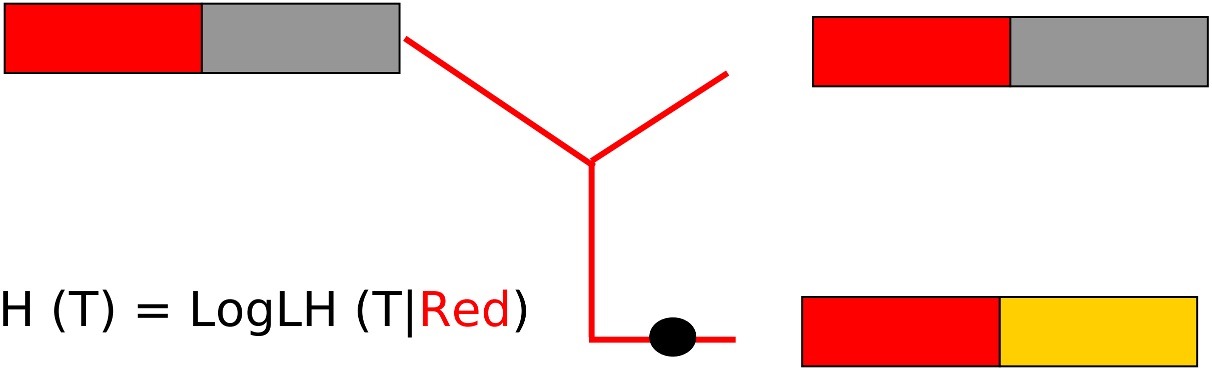
A Multi-Gene Model



A Multi-Gene Model



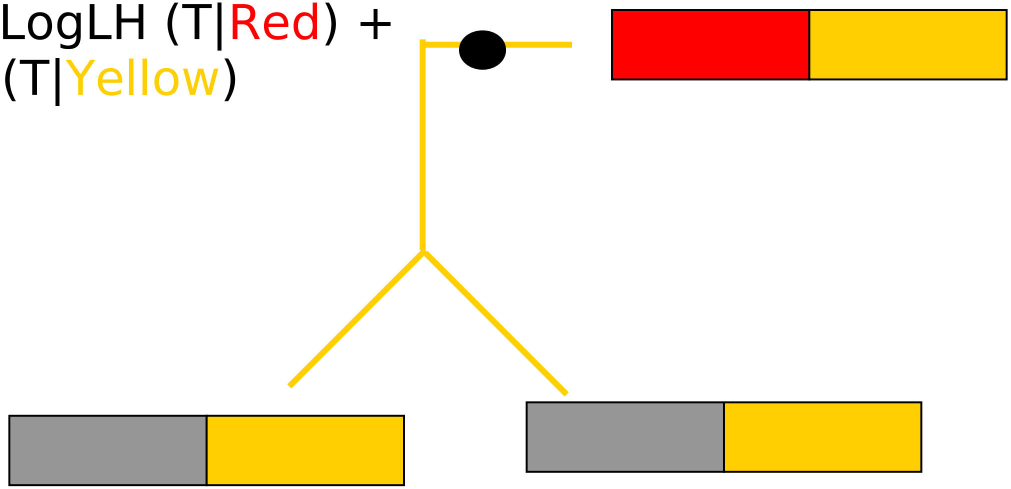
A Multi-Gene Model



LogLH (T) = LogLH (T|Red)

A Multi-Gene Model

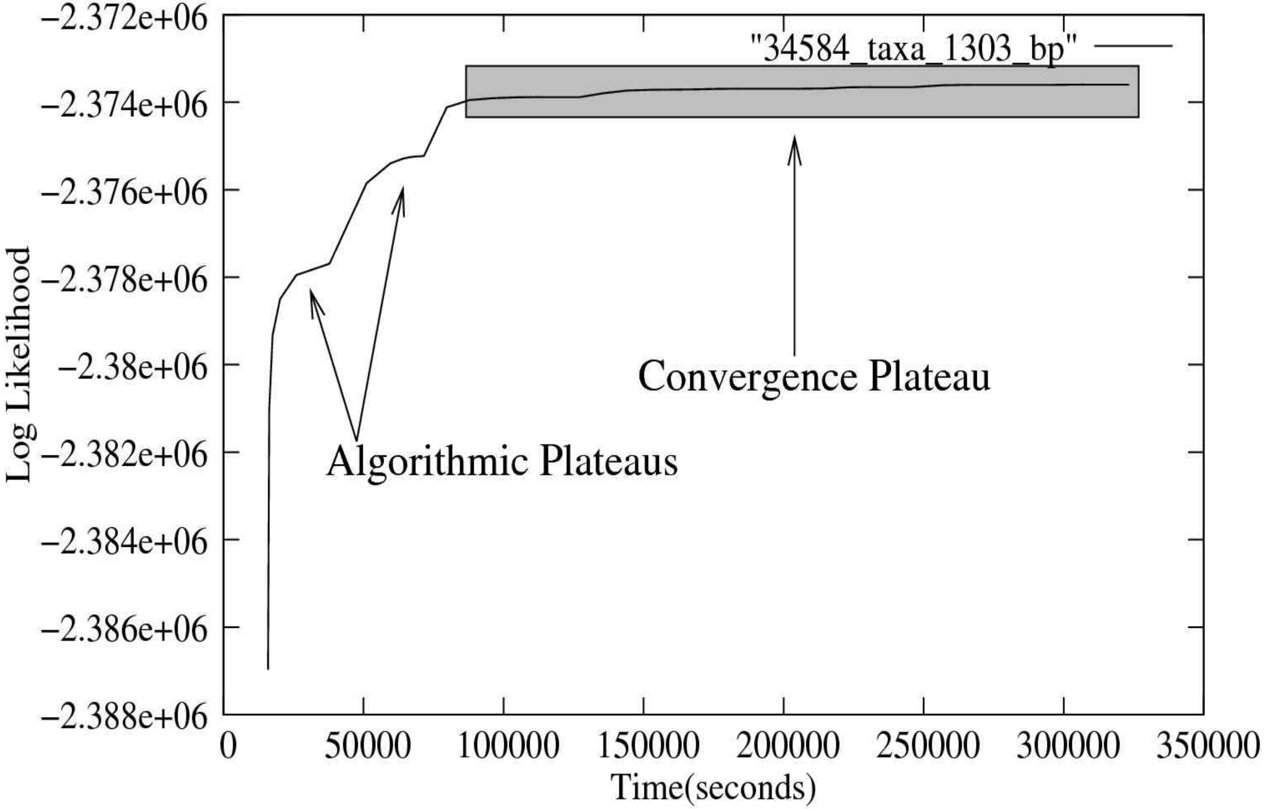
LogLH (T) = LogLH (T|Red) + LogLH(T|Yellow)



How to do Searches (SPRs)?

- Proof of concept SPRs
- Test dataset:
  - 2177 taxa
  - 68 partitions
  - Gappyness 89.53%
  - Memory 9.0 -> 1.1 GB
  - Fast SPRs: speedup 11.8
  - Slow SPRs: speedup 10.6
- **Deliverable 2010**

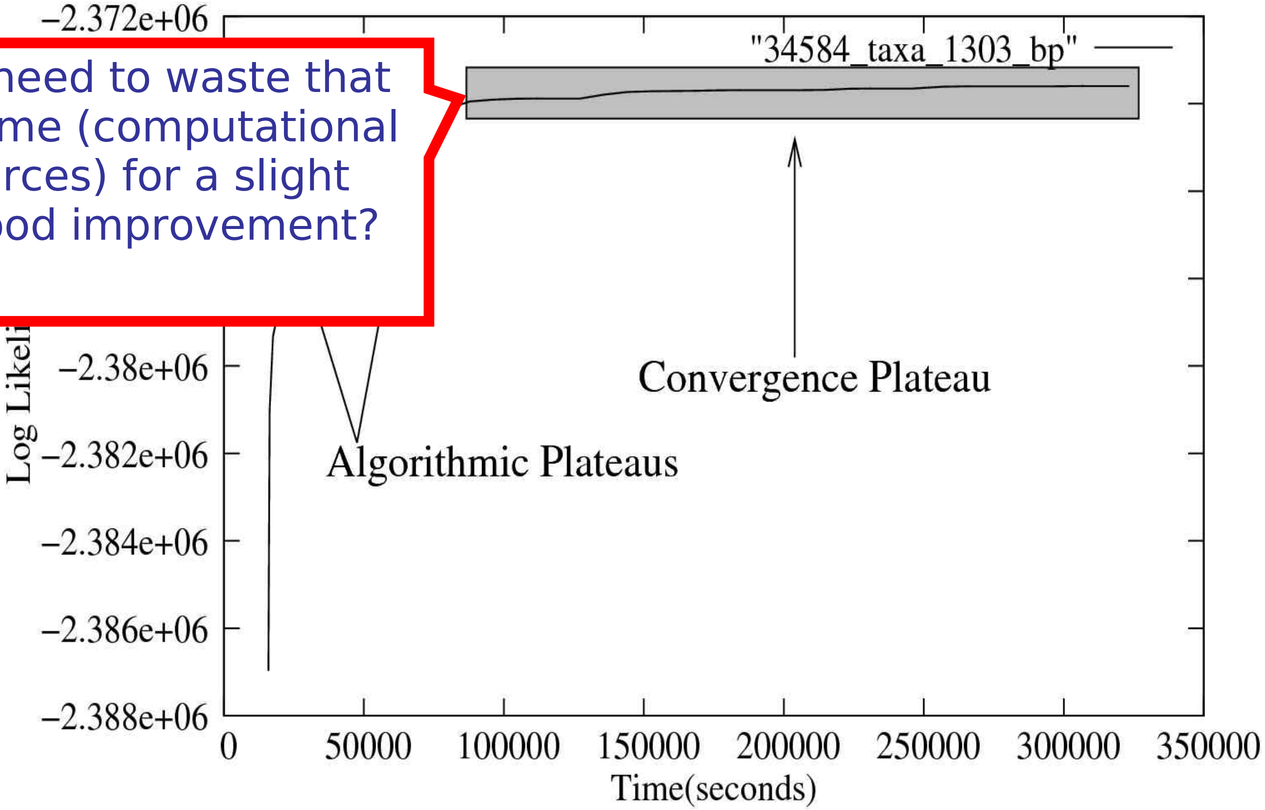
ML Search Convergence



Convergence Plateau

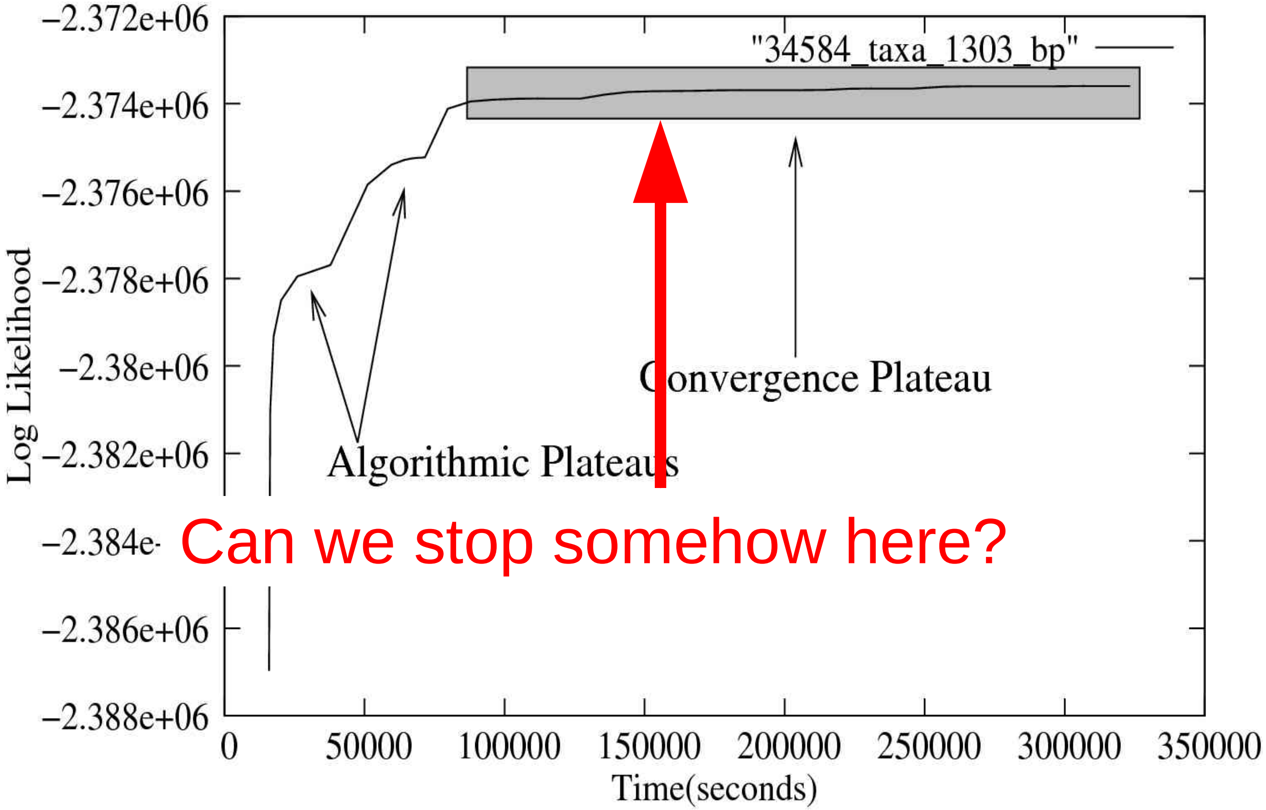
Algorithmic Plateaus

ML Search Convergence



Do we need to waste that much time (computational resources) for a slight likelihood improvement?

ML Search Convergence



Can we stop somehow here?

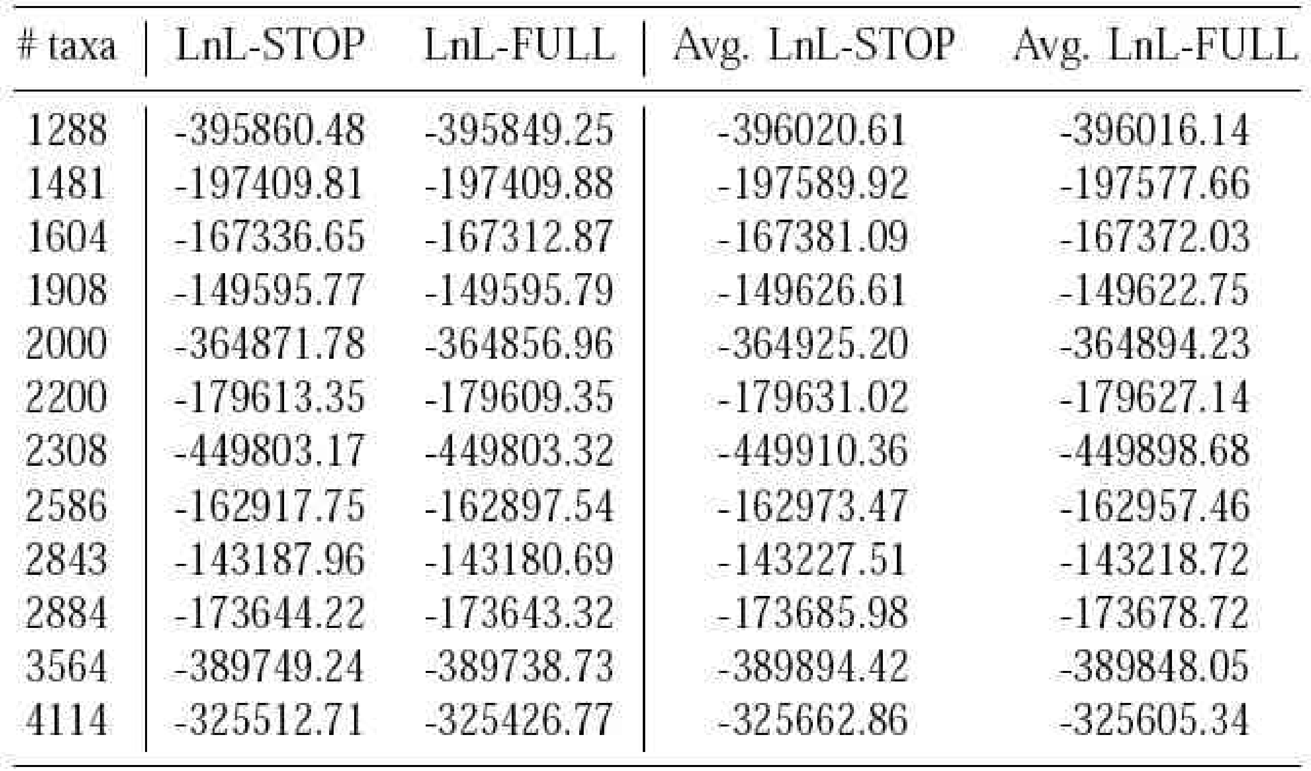
Stopping Rule

- Stop ML search if the trees generated by two successive SPR cycles have a RF distance < 1%
- Stephen's dataset
  - 9,028 bp x 37,840 taxa
  - 1 tree search, 16 threads
  - Fast: 75 hours   LnL = -5575995
  - Slow: 207 hours LnL = -5575582
- Allows for reconstructing 3x more trees with the same amount of computational resources
- Better sampling of rough likelihood surfaces :-)

More Results

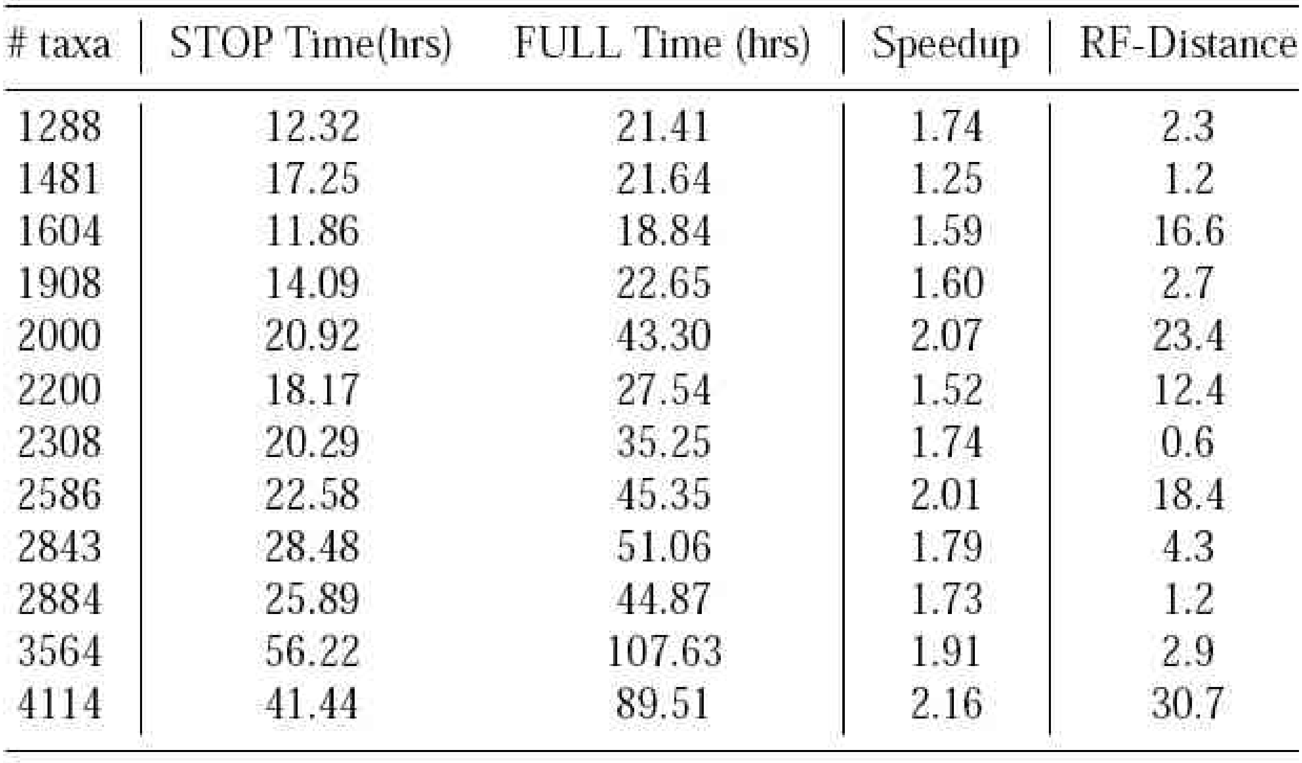
- Thorough test:
- 12 single-gene datasets 1,288 – 4,144 taxa
- Computed 40 ML trees per dataset with and without convergence criterion
- **Deliverable: Already available**

Likelihoods



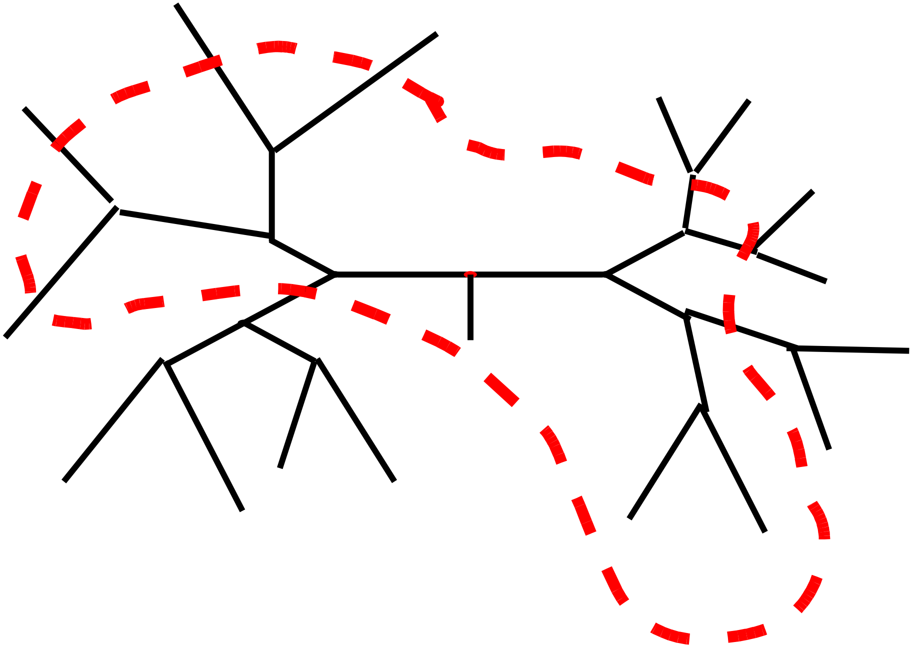
# taxa	LnL-STOP	LnL-FULL	Avg. LnL-STOP	Avg. LnL-FULL
1288	-395860.48	-395849.25	-396020.61	-396016.14
1481	-197409.81	-197409.88	-197589.92	-197577.66
1604	-167336.65	-167312.87	-167381.09	-167372.03
1908	-149595.77	-149595.79	-149626.61	-149622.75
2000	-364871.78	-364856.96	-364925.20	-364894.23
2200	-179613.35	-179609.35	-179631.02	-179627.14
2308	-449803.17	-449803.32	-449910.36	-449898.68
2586	-162917.75	-162897.54	-162973.47	-162957.46
2843	-143187.96	-143180.69	-143227.51	-143218.72
2884	-173644.22	-173643.32	-173685.98	-173678.72
3564	-389749.24	-389738.73	-389894.42	-389848.05
4114	-325512.71	-325426.77	-325662.86	-325605.34

Speedups



# taxa	STOP Time(hrs)	FULL Time (hrs)	Speedup	RF-Distance
1288	12.32	21.41	1.74	2.3
1481	17.25	21.64	1.25	1.2
1604	11.86	18.84	1.59	16.6
1908	14.09	22.65	1.60	2.7
2000	20.92	43.30	2.07	23.4
2200	18.17	27.54	1.52	12.4
2308	20.29	35.25	1.74	0.6
2586	22.58	45.35	2.01	18.4
2843	28.48	51.06	1.79	4.3
2884	25.89	44.87	1.73	1.2
3564	56.22	107.63	1.91	2.9
4114	41.44	89.51	2.16	30.7

Zoom-in Zoom-out



Dimensionality reduction challenge:
Where and how to cut off?

RAxML v7.2.4

- New stuff:
  - Methods for accurate fossil placement
  - Multi-state characters
  - Parallelization & optimization of operations on trees (consensi etc) already factor 40 via sequential optimization
  - Hybrid MPI/Pthreads version
- Exelixis Rapid Research Dissemination Reports: http://wwwkramer.in.tum.de/exelixis/publications.html

The Non-ML World

- Pablo Goloboff TNT: 70K taxa *Cladistics*, 2009.
- NJ tree building programs (under integration):
  - Exact NJ: **Ninja**, *Travis J. Wheeler*, 100K taxa
  - Approximate NJ & ML: **FastTree 2.0**, *Morgan Price*, 200K taxa

Acknowledgments

Simon Berger, TUM

Nikos Alachiotis, TUM

Stephen Smith, NESCENT

Michael Ott, TUM

Andre Aberer, TUM

Nick Pattengale, UNM

Wayne Pfeiffer, SDSC

Thank you for your Attention !

Sfakia, Crete, Greece, August 2009