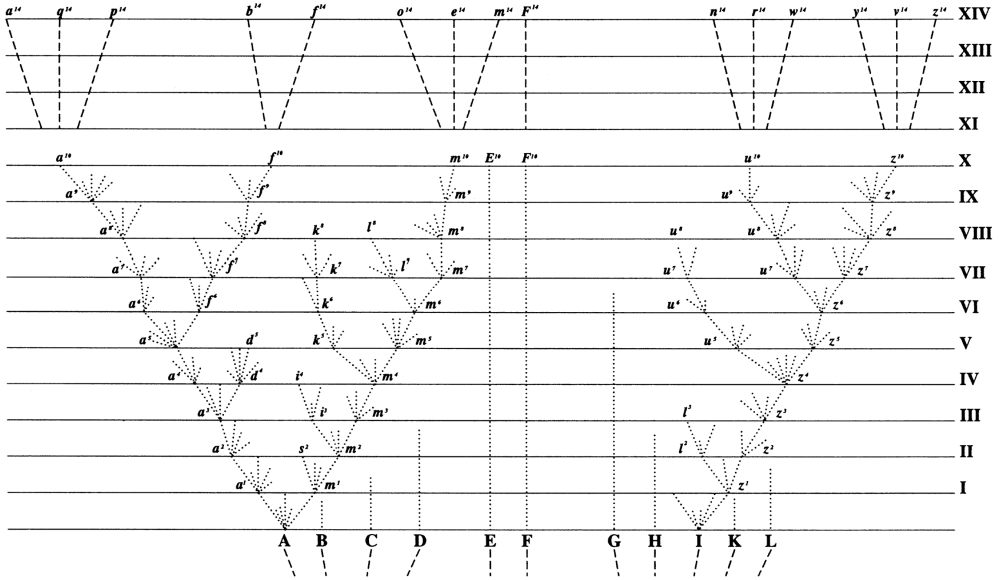
# IQPNNI: Moving fast through tree space and stopping in time

Vinh Le Sy[1] & Arndt von Haeseler[1,2]

[1]Neumann Institute for Computing, FZ-Jülich, Germany

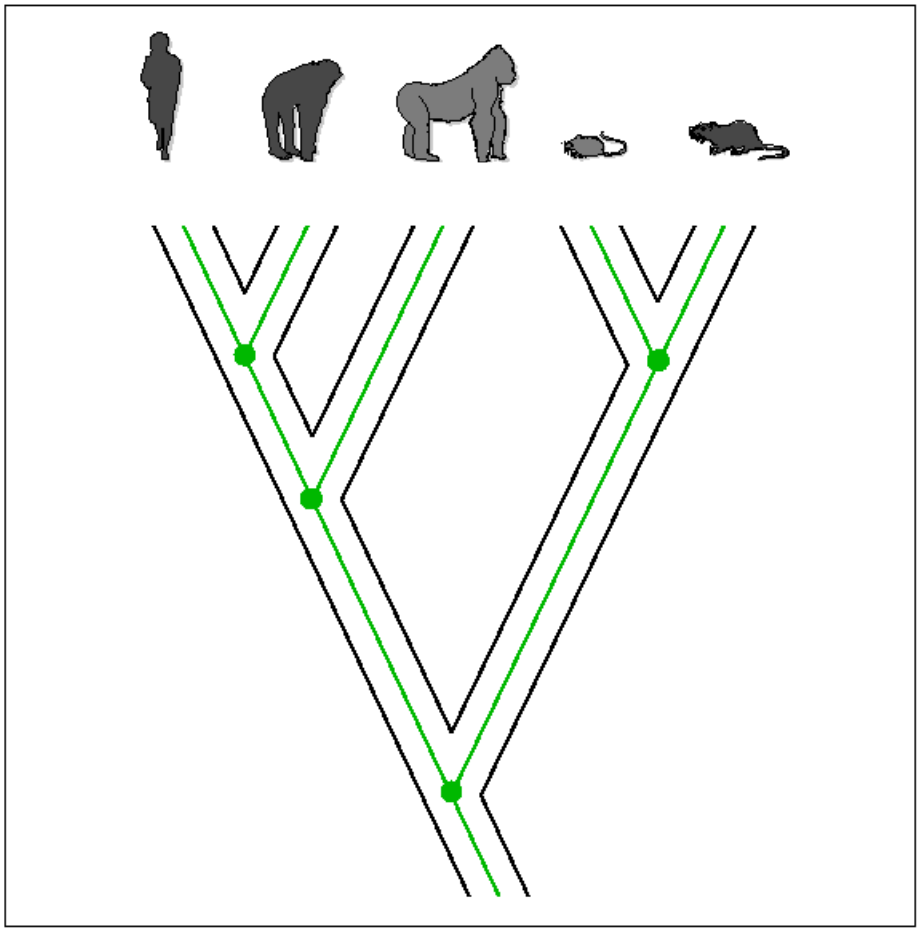
[2]Heinrich Heine University, Düsseldorf, Germany

# Charles Darwin: On the Origin of Species



# Multiple Sequence Alignment

| | Site 0 | Site 1 | Site 2 | Site 3 | Site 4 | Site 5 | Site 6 | Site 7 | Site 8 | ... | Site N–2 | Site N–1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 0 | A | T | A | A | A | A | G | C | T | ... | A | T |
| Sequence 1 | C | G | A | G | G | C | G | C | C | ... | T | G |
| Sequence 2 | A | C | C | T | C | T | G | C | G | ... | A | G |
| Sequence 3 | A | C | G | G | G | T | T | C | A | ... | A | T |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Sequence S–1 | T | C | G | A | G | T | A | C | T | ... | A | C |



# Data and Trees

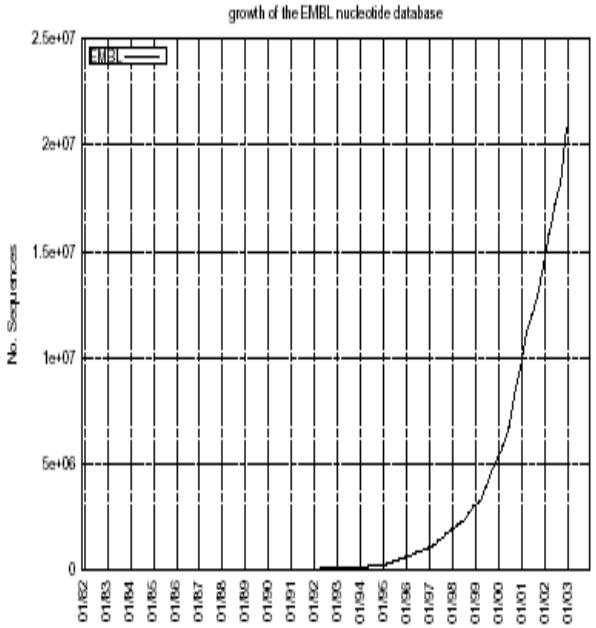
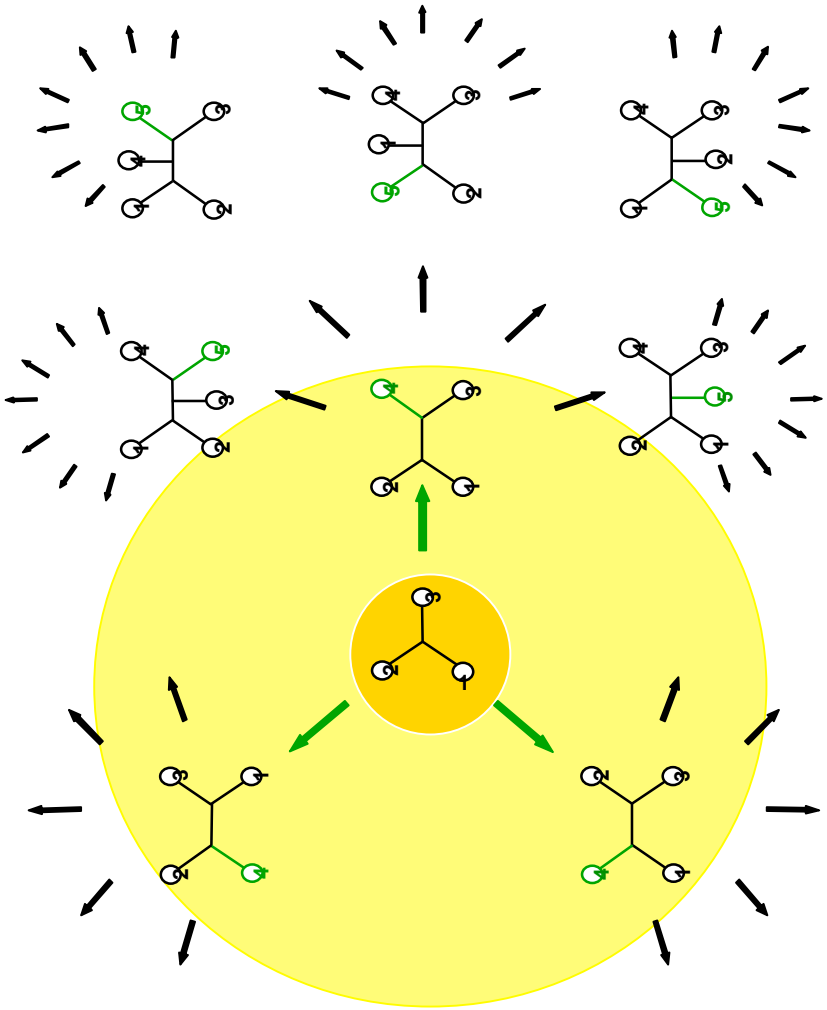



Figure 1.1: Public nucleotide databases like the EMBL database show exponential growth, doubling their content ever



# Molecular Phylogenetics

Reconstruct a phylogenetic tree based on DNA or amino acid sequences.

Tree reconstruction programs (sequential)

1. MOLPHY Adachi & Hasegawa (1992)
2. PHYLIP Felsenstein (1993)
3. MEGA Kumar & Nei (1994)
4. PAUP Swofford et al. (1996)
5. PUZZLE Strimmer & von Haeseler (1996)
6. PAML Yang (1997)

# Programs for large number of sequences

- Olsen et al. (1992) fastDNAml
- Gascuel (1997) BIONJ
- Huson et al. (1999) Disc-covering
- Bruno et al. (2000) Weighbor
- Lemmon & Milinkovitch (2002) MetaPIGA
- Guindon & Gascuel (2002) PHYML
- Schmidt et al. (2002) parallel PUZZLE
- Vos (2003) likelihood ratched

# Heuristic searches through tree space

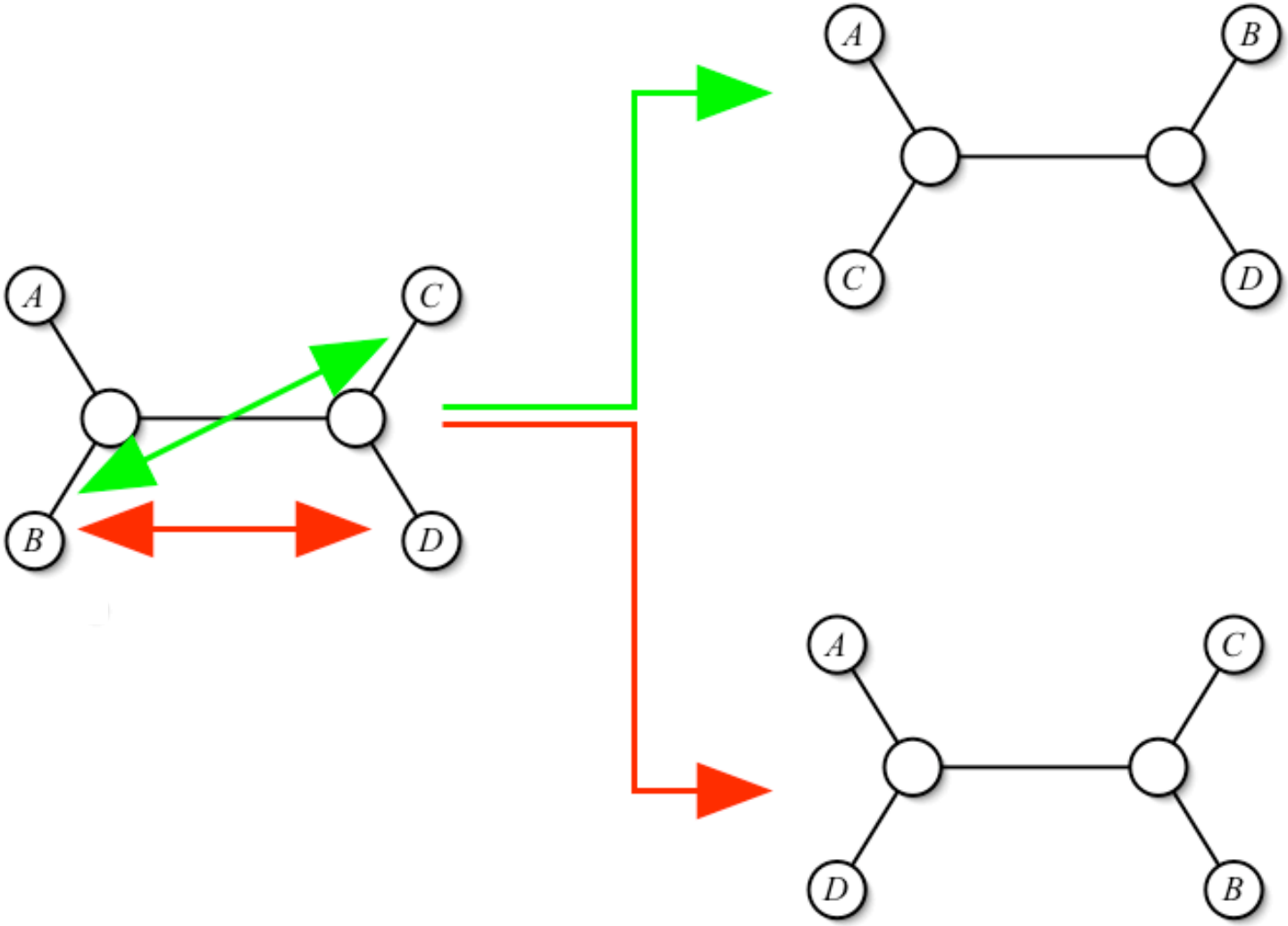
- Branch swapping:

  Nearest neighbour interchange (NNI)

- Sequential addition of single sequences:

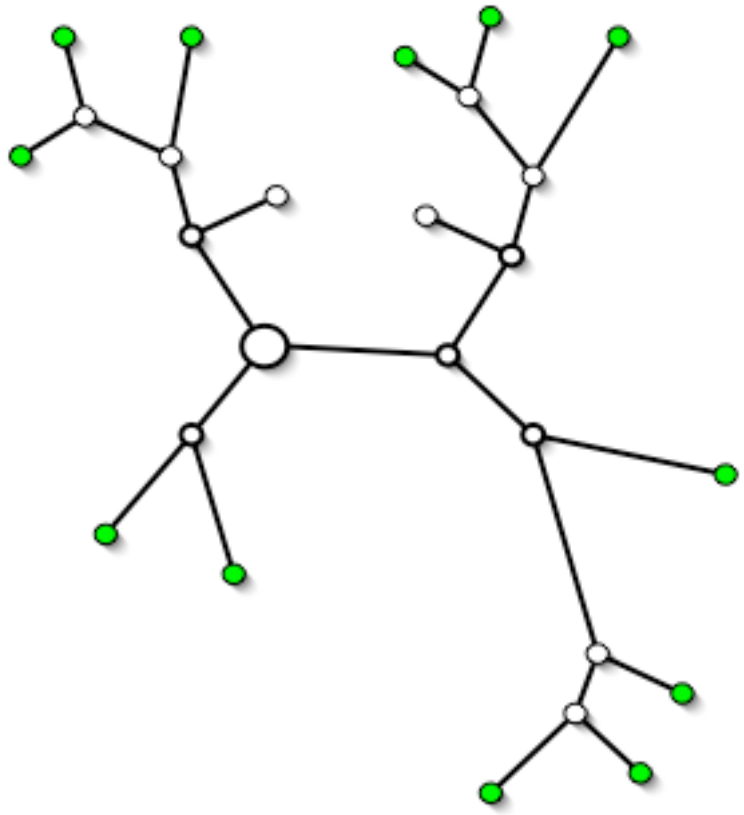
  TREE-PUZZLE  reconstruct trees from their buildig blocks: the quartets.

# Branch Swapping: NNI

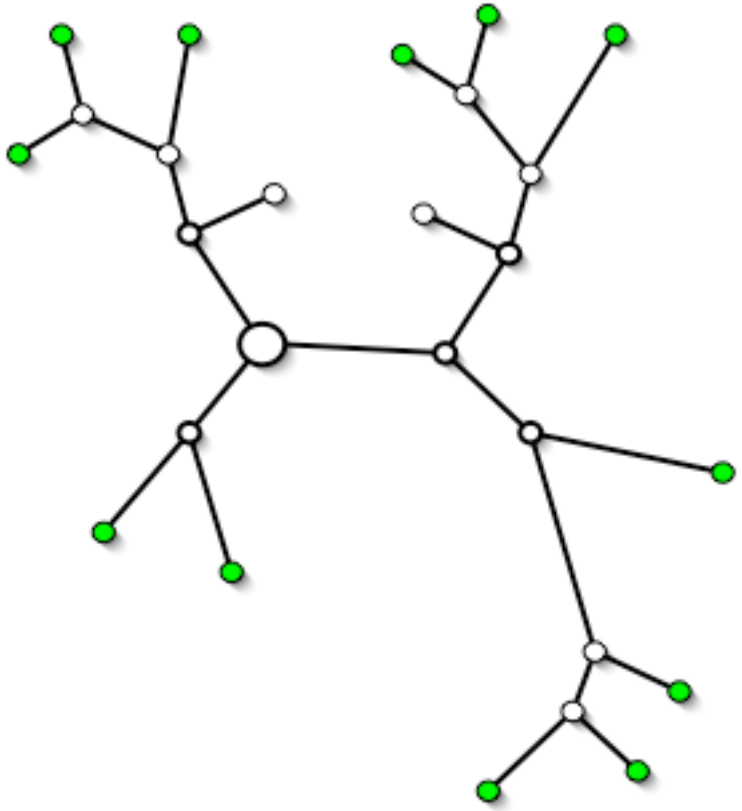


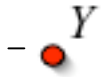
# Sequential Addition: Tree Puzzle

partial tree

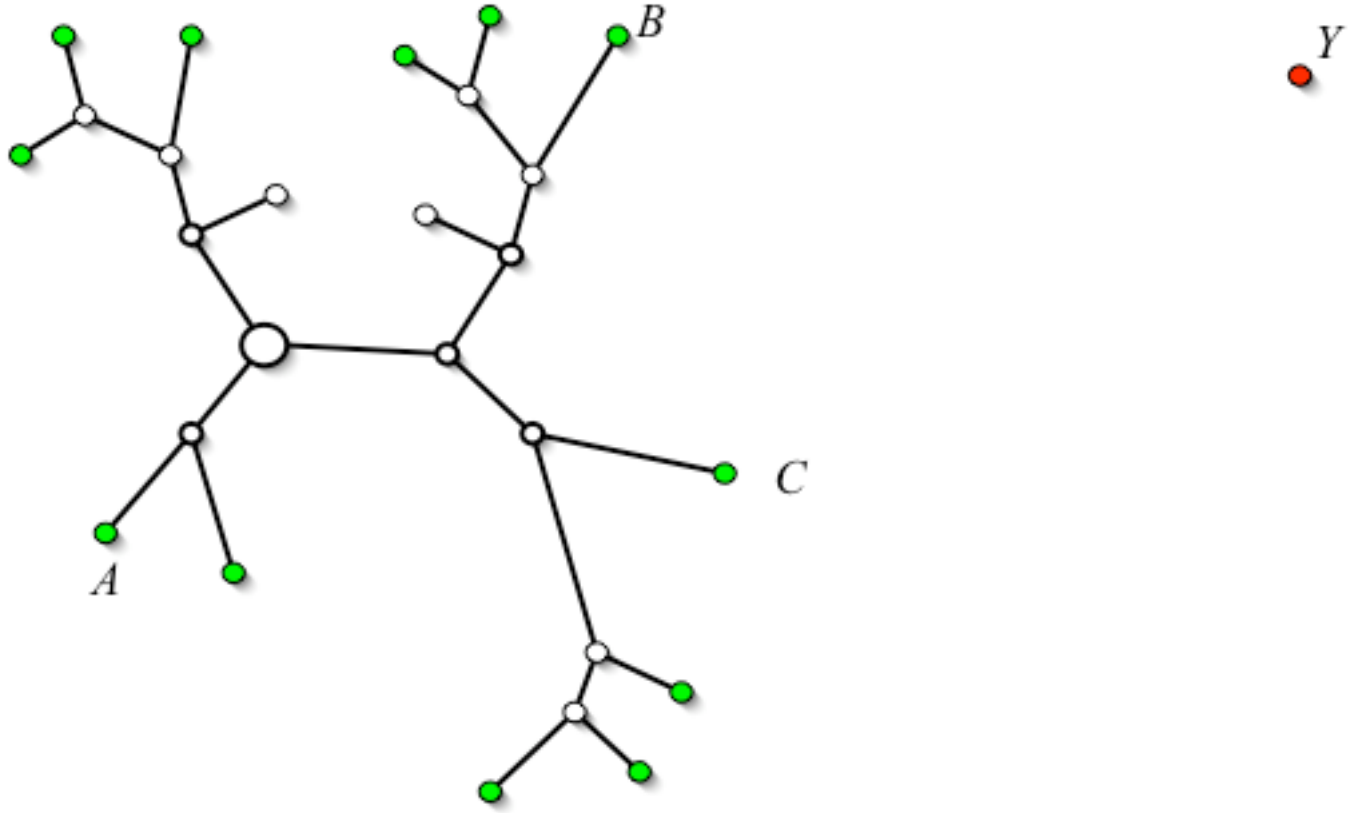


# Tree Puzzle

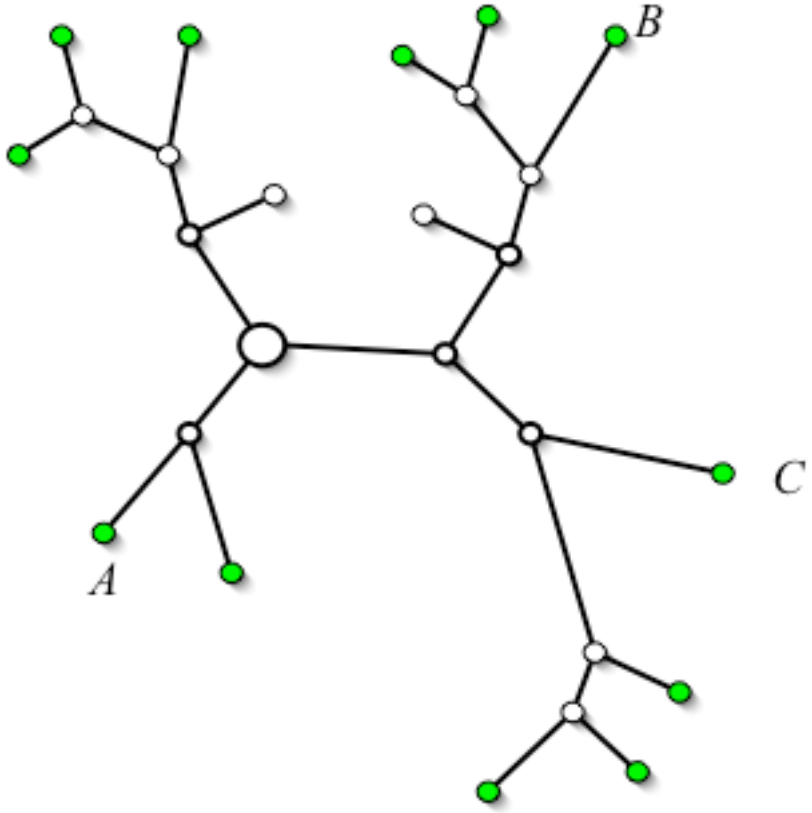




# Tree Puzzle



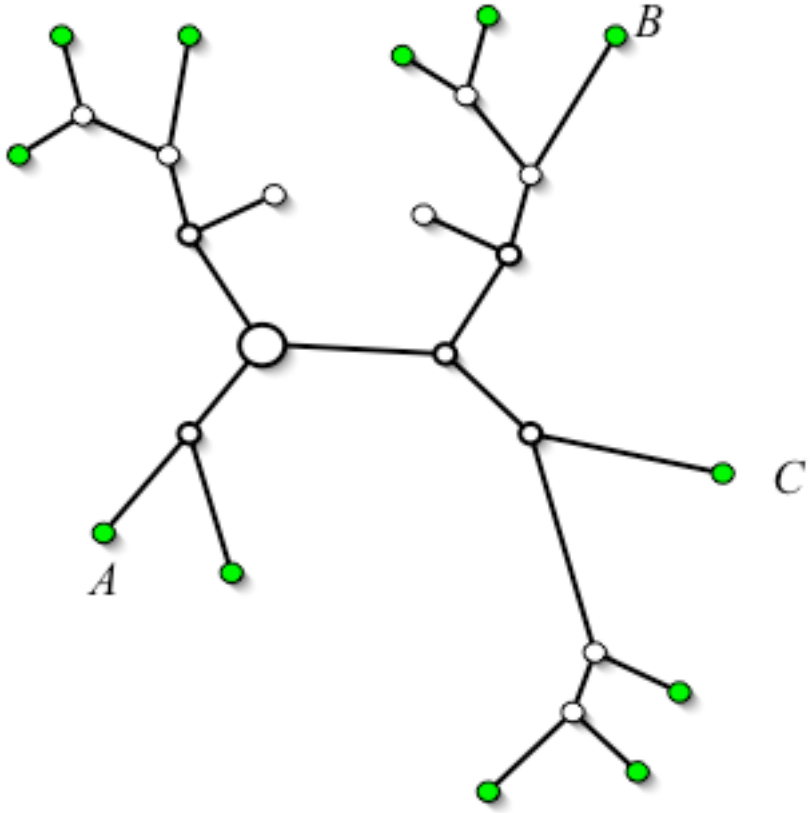
# Tree Puzzle

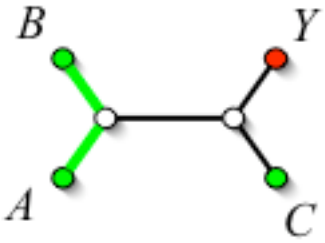




optimal four seq. tree

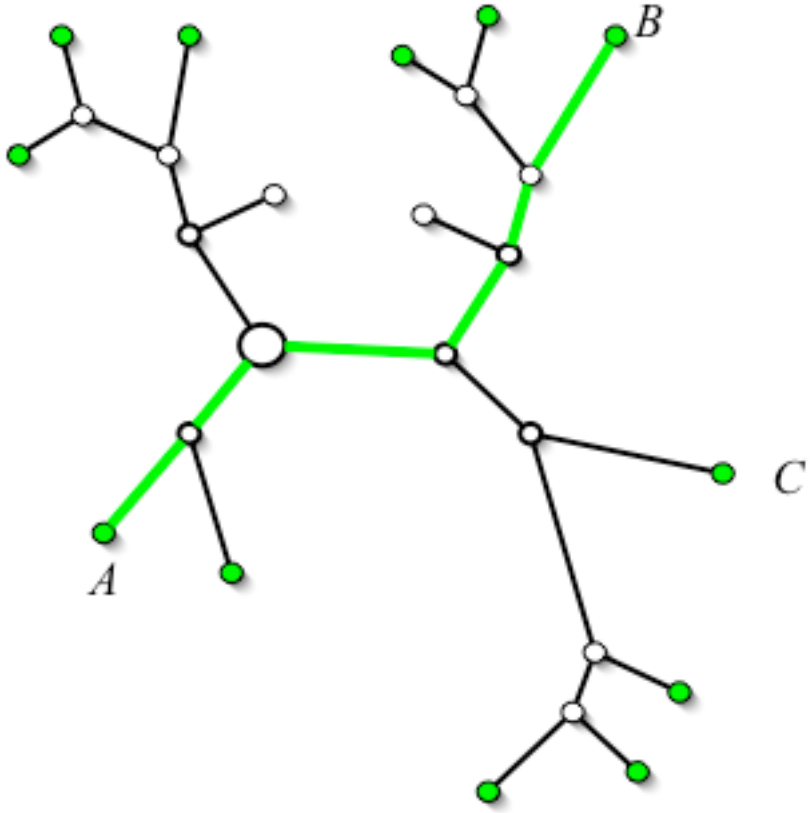
# Tree Puzzle

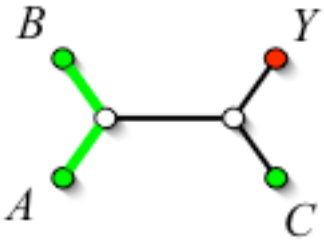



optimal four seq. tree

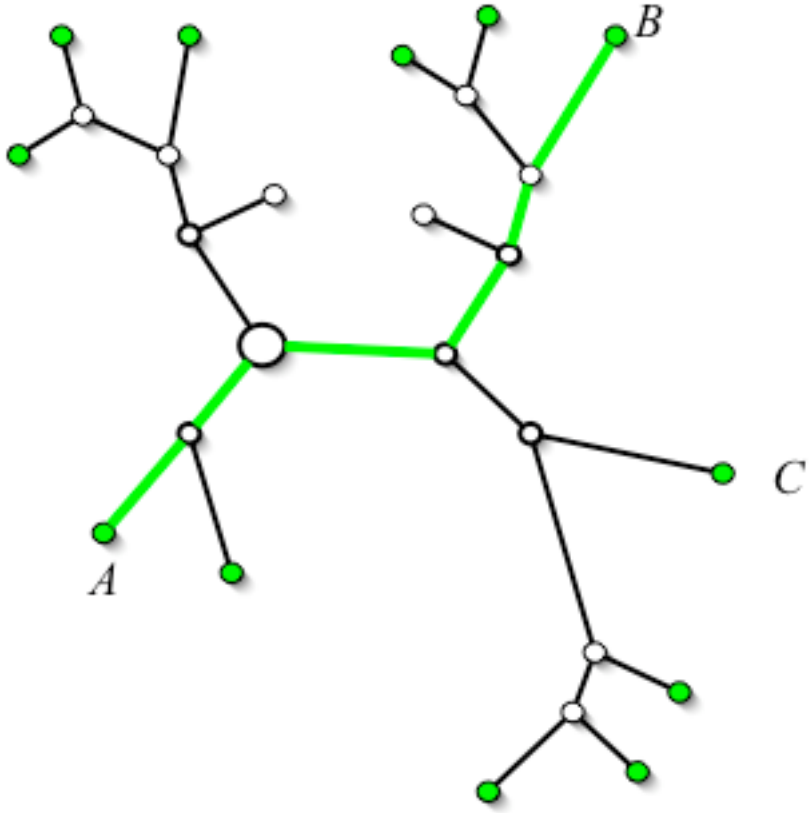
# Tree Puzzle

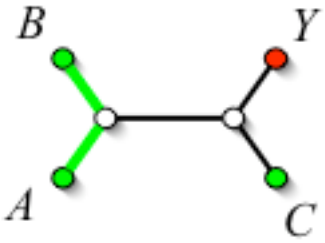



optimal four seq. tree

# Tree Puzzle



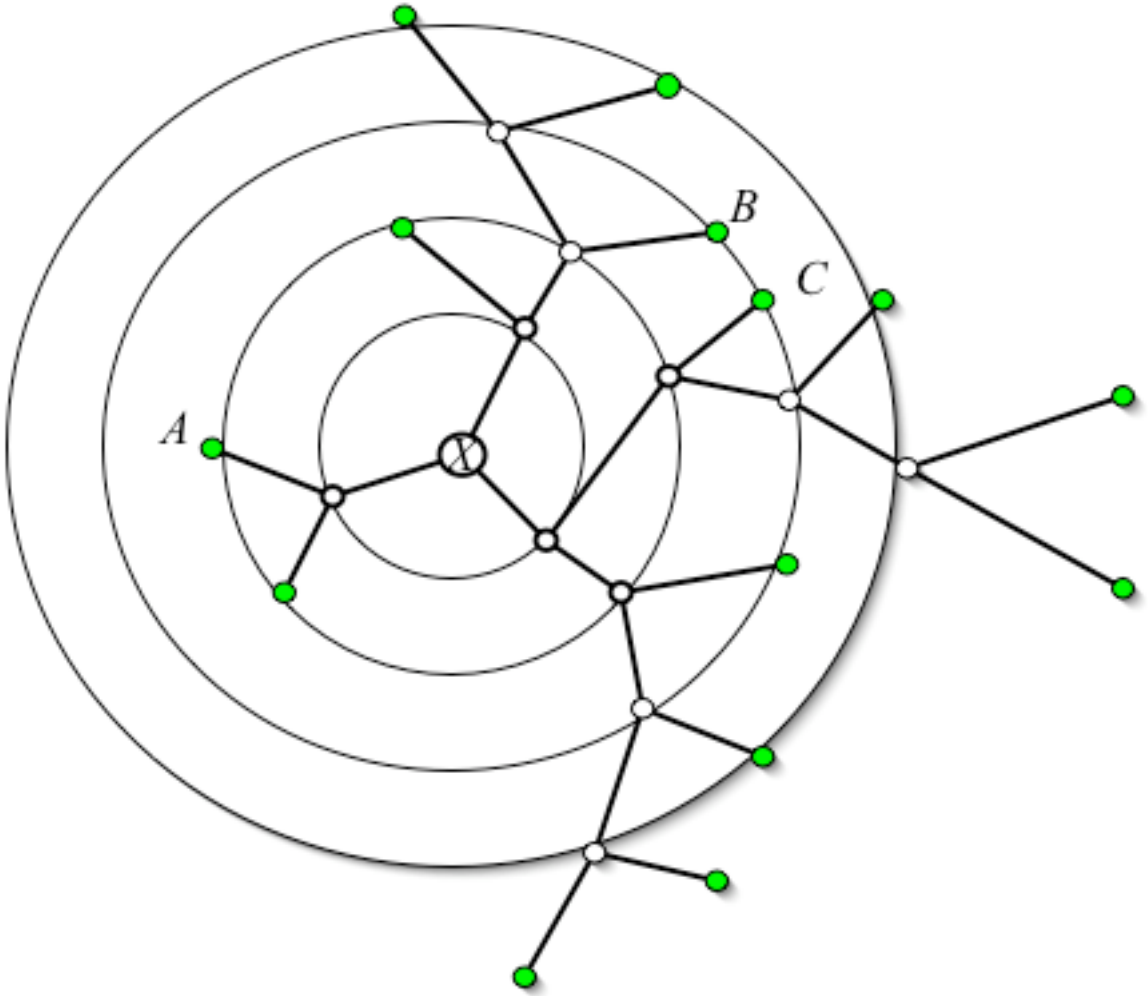




Repeat this for all quartets and place Y on the branch with the smallest penalty.

# Complexity

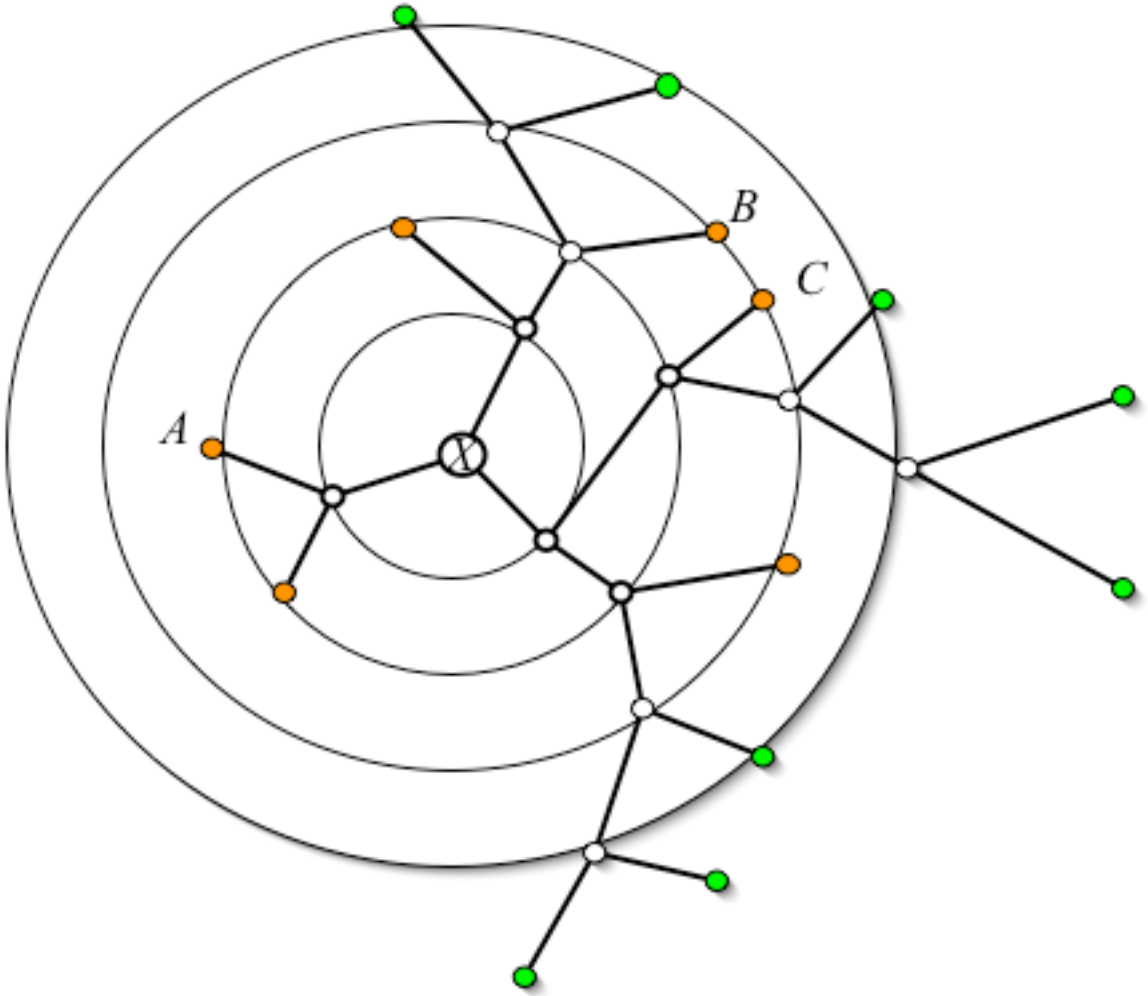
$$\sum_{k=4}^{n}\binom{k}{3} = \frac{n^4}{24} - \frac{n^3}{12} - \frac{n^2}{24} + \frac{n}{12} - 1 = O(n^4)$$

# Important Quartet Puzzle



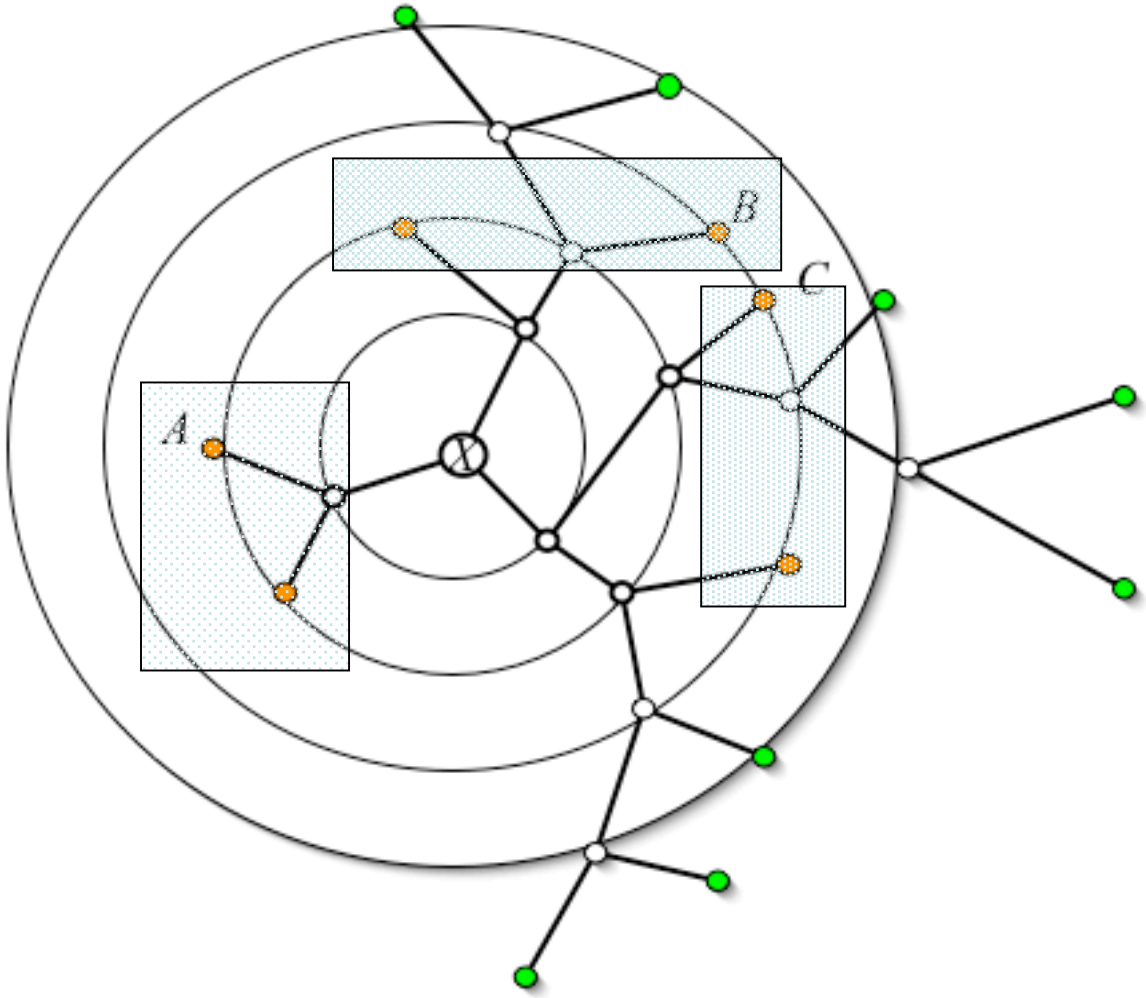

concentric circles show the distance of  nodes to node *X*

# Important Quartet Puzzle




Compute the $k$=2 nearest leaves for each subtree emerging from $X$

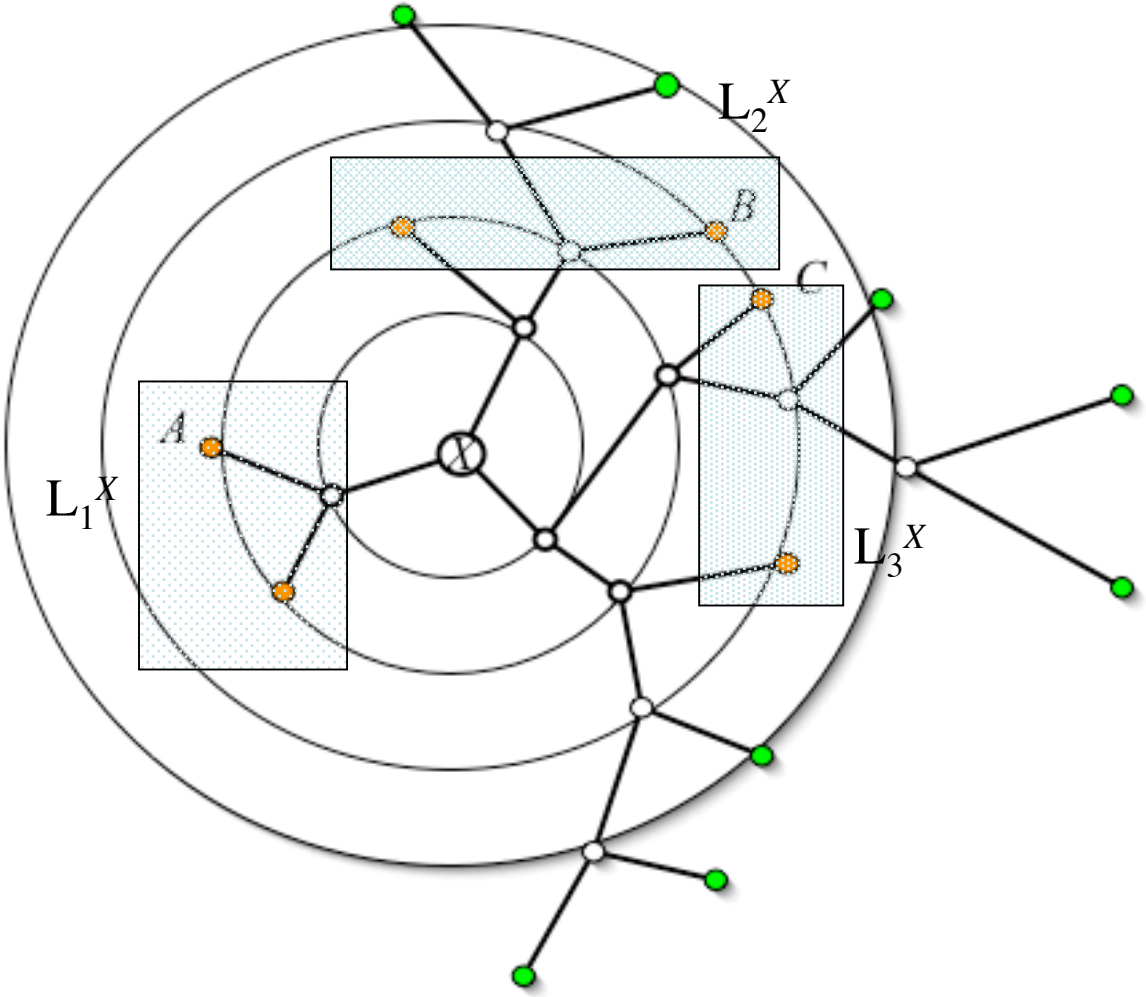
# Important Quartet Puzzle




representative
leaf sets
$L_1^X$, $L_2^X$, $L_3^X$

# Important Quartet Puzzle

Compute quartet trees for each triple from $L_1^X$ ,$L_2^X$, $L_3^X$, and Y

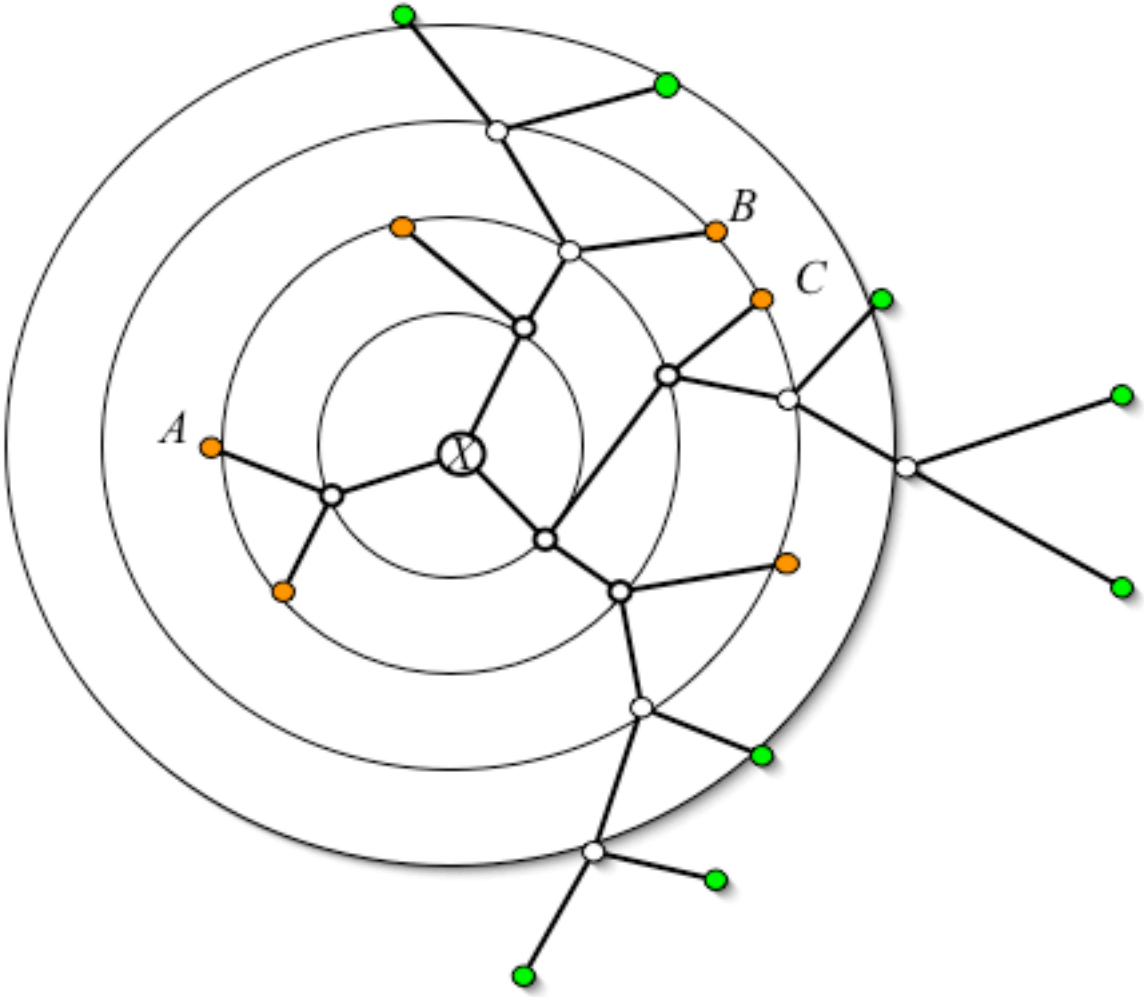




optimal four seq. tree

# Important Quartet Puzzle

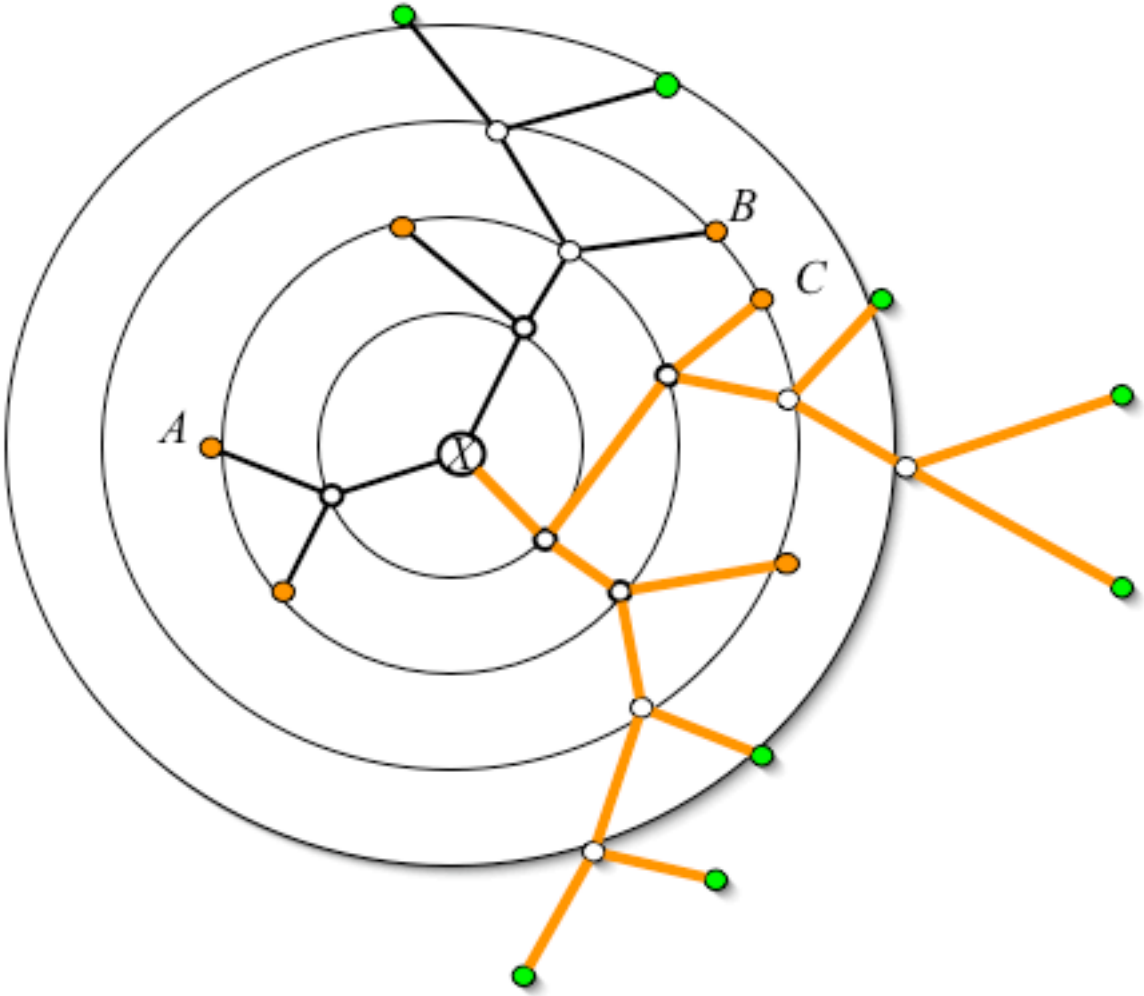
Where to place *Y*?





optimal four seq. tree

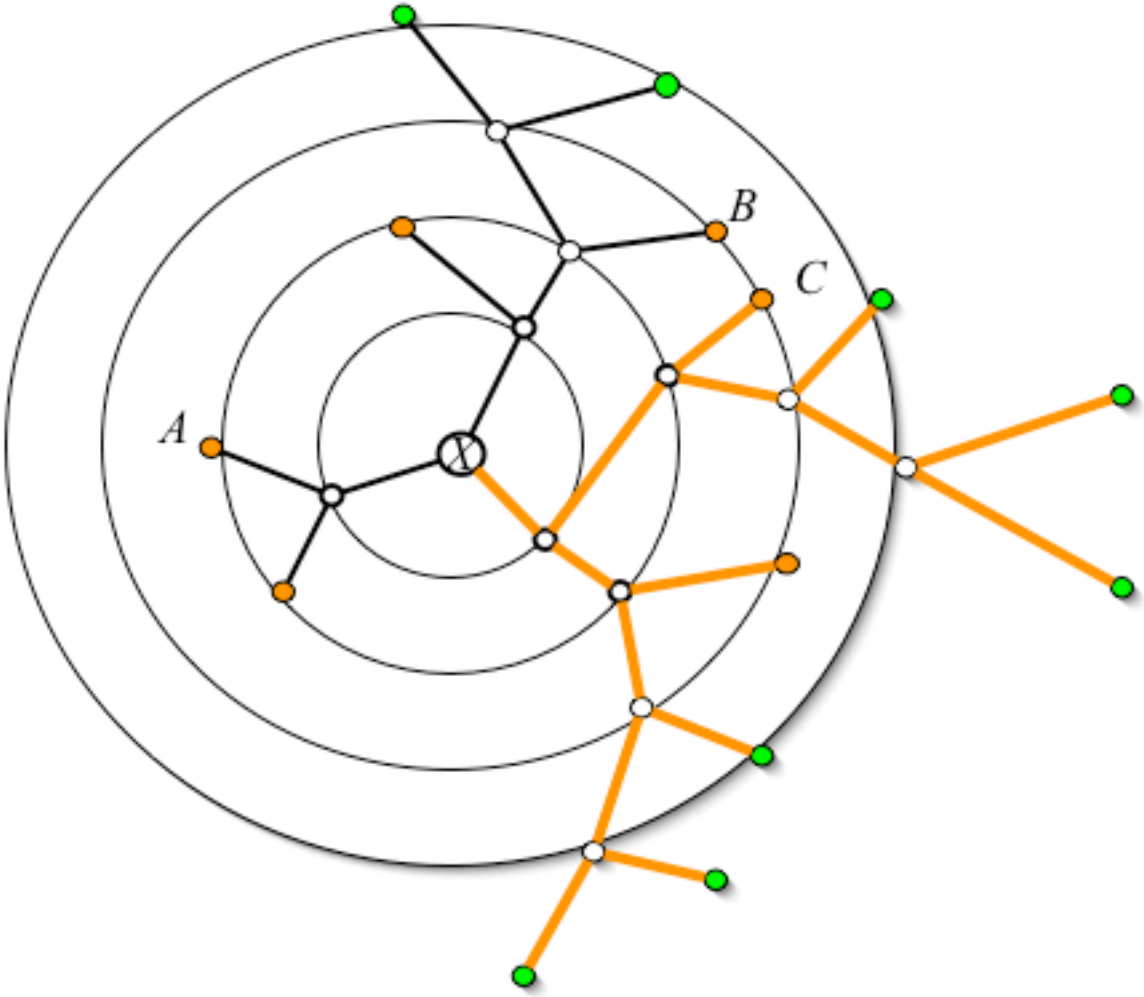
# Important Quartet Puzzle





optimal four seq. tree

# Important Quartet Puzzle


Repeat this for all quartets from the representative leaf sets and for all internal nodes.

Place *Y* at the branch with highest score.

# Complexity

$n$ internal nodes with $k^3$ quartets per node, thus $O(n^2)$ is the complexity.

# Combining IQP and NNI

1. **initial step:** Build a starting tree using BIONJ. Then perform NNI until no further improvement of the likelihood is found.

2. **optimization step:** Delete with probability *p* each leaf from the current tree. Re-insert the deleted leaves by applying IQP. Optimize the resulting tree via NNI. The resulting tree is called intermediate tree.

3. **comparative step:** If the log-likelihood of the intermediate tree is higher than the current best tree, take the intermediate tree as the current best tree.

4. **stop-criterion:** if the number of optimization steps exceeds a pre-defined number *M* stop, otherwise go to 2.

# Does it work?

## Accuracy: Ability to reconstruct a simulated tree

Simulations: 3,000 trees with 30 sequences,
500 bp, Kimura 2P model.
trees drawn from a Yule-Harding distribution.
branch lengths from an exponential with
mean 0.03, 0.06, and 0.15.

# Computing time (min) for 1000 sequences

| b.p. | Weighbor | PHYML | IQPNNI per tree |
|---|---|---|---|
| 500 | 190.0 | 6.5 | 2.7 |
| 1000 | 190.0 | 13.5 | 4.3 |
| 2000 | 172.0 | 19.0 | 6.3 |

# real time (min) for 1000 sequences

| b.p. | Weighbor | PHYML | IQPNNI total |
|---|---|---|---|
| 500 | 190.0 | 6.5 | 270.0 |
| 1000 | 190.0 | 13.5 | 430.0 |
| 2000 | 172.0 | 19.0 | 630.0 |

# Biological data

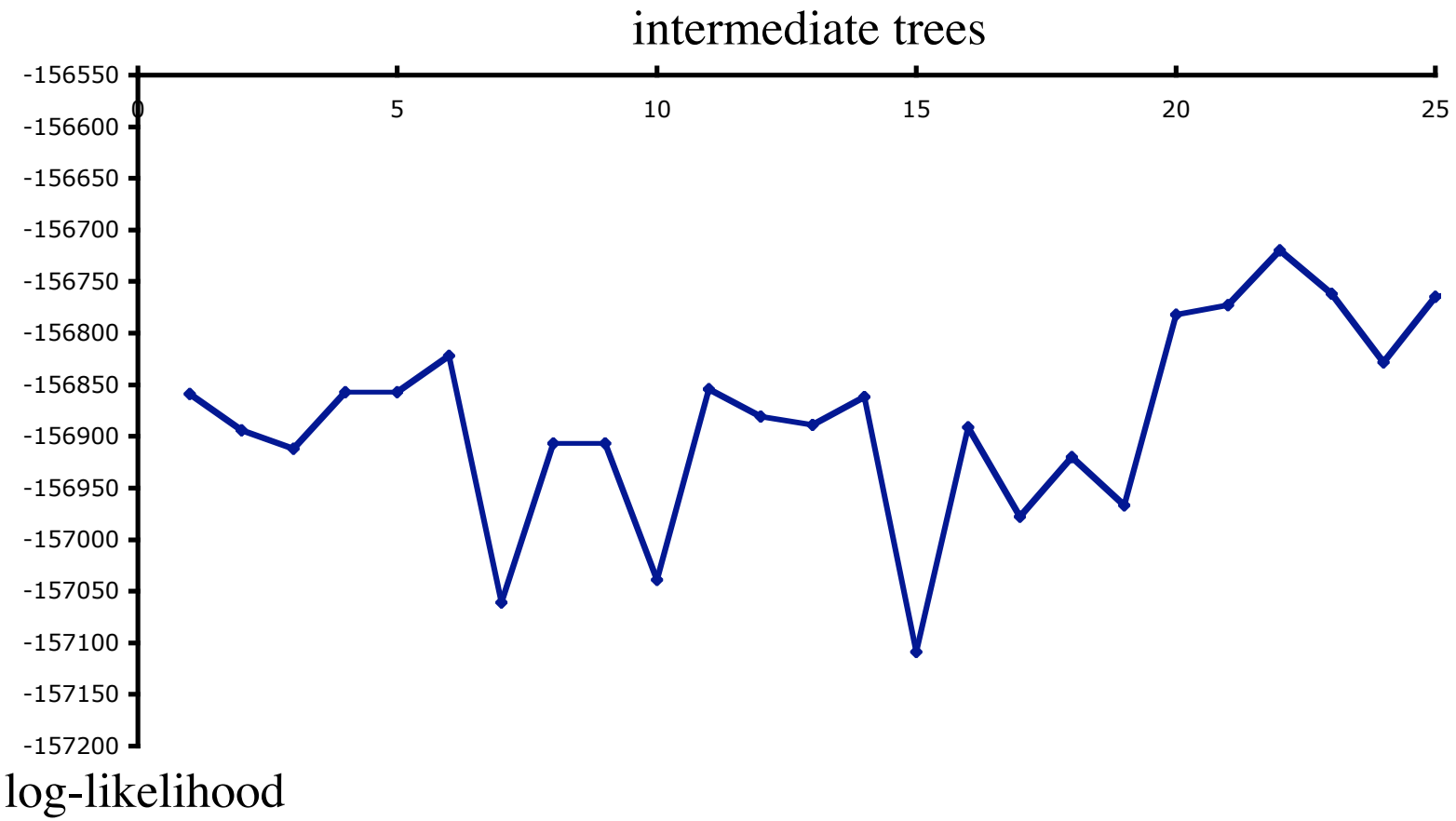
Two data sets

1. ssu-rRNA 218 sequences, 4182 bp long
2. rbcl-genes 500 sequences, 1398 bp long

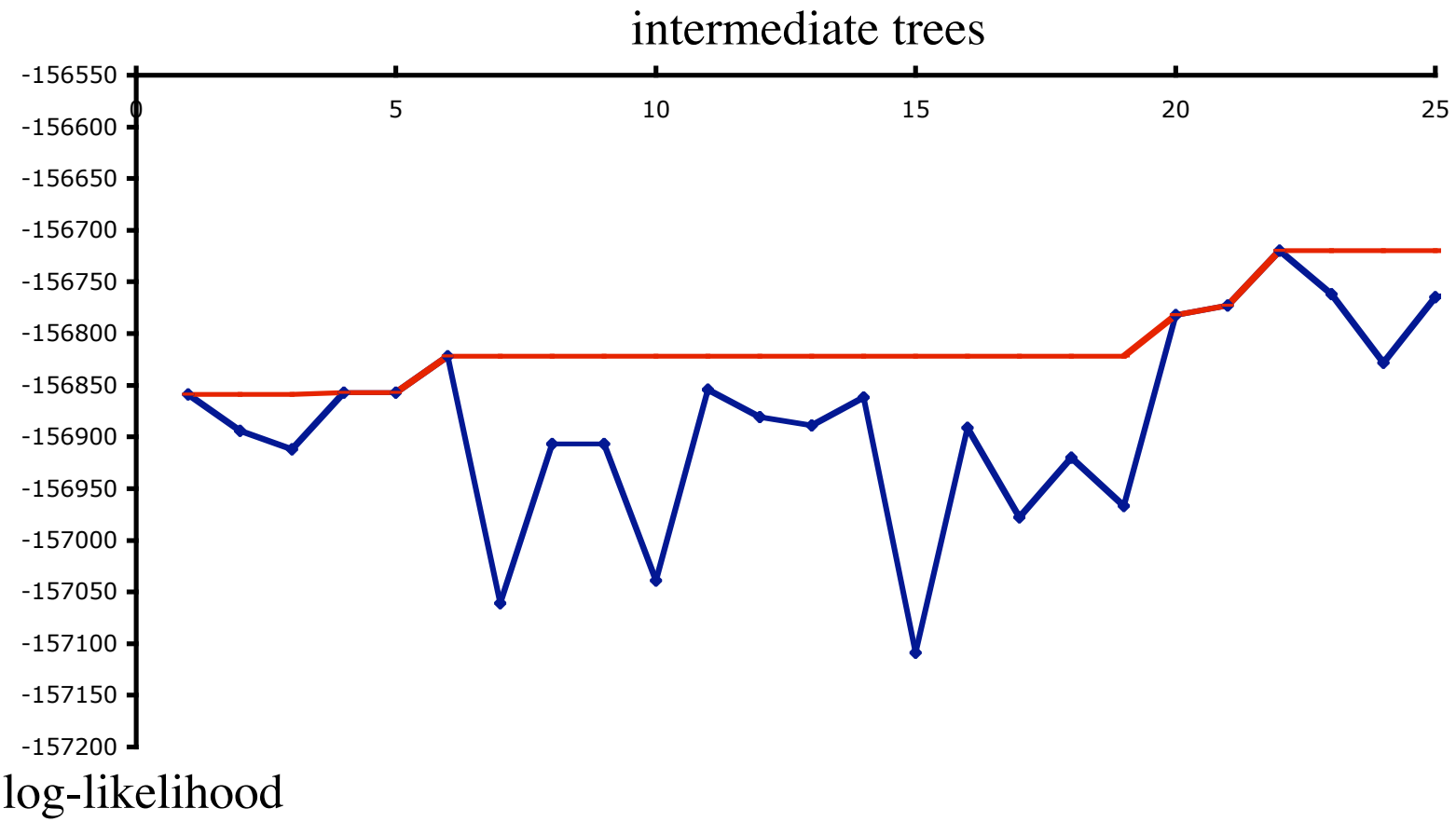
# Log-likelihood for real data

| | | PHYML | MetaPIGA | IQPNNI |
|---|---|---|---|---|
| gene | Number of sequences | Log likelihood | | |
| ssu rRNA | 218 | -156,895 | -156,715 | -156,604 |
| rbcl | 500 | -100,191 | -100,080 | -100,011 |

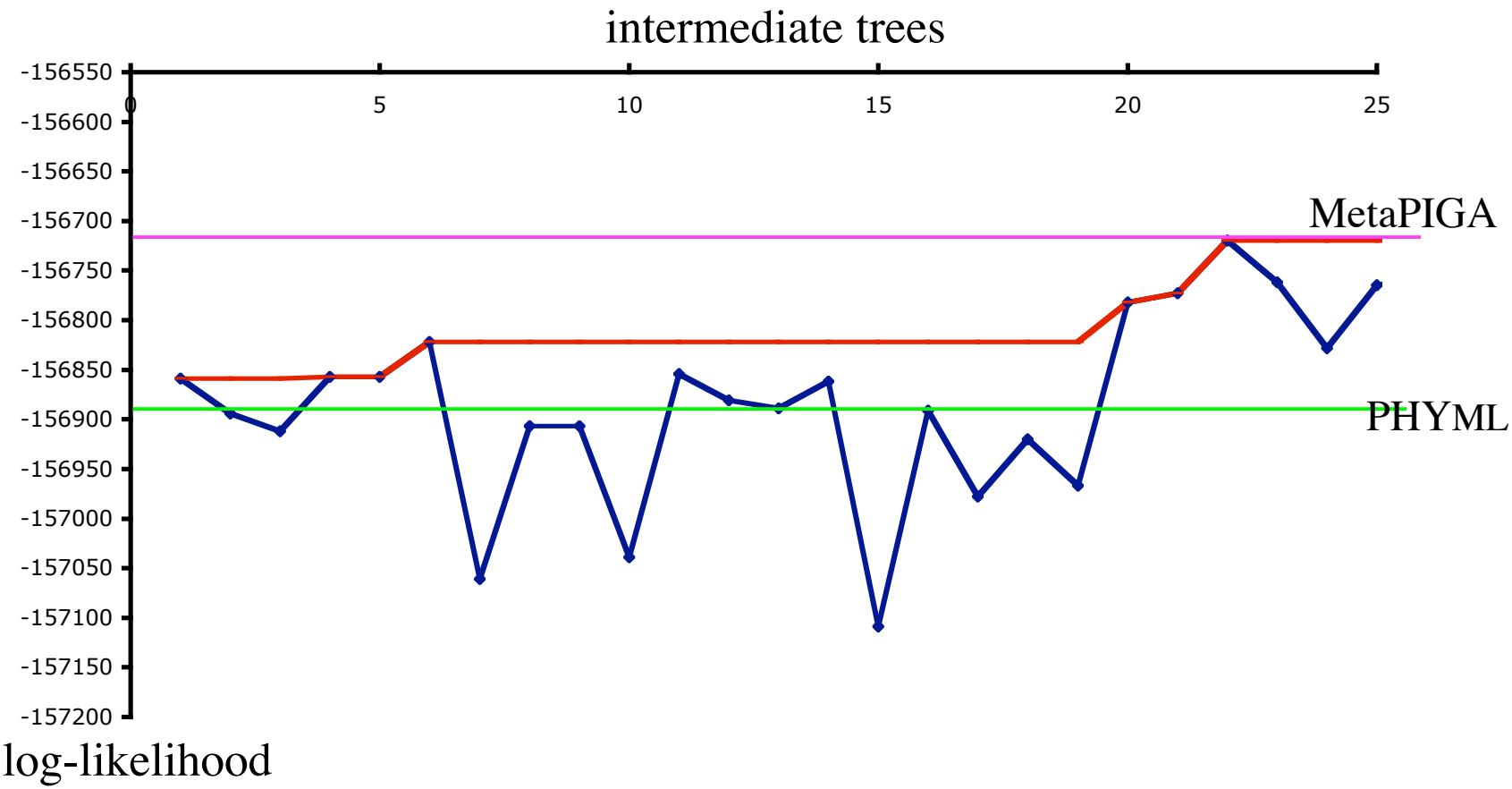
# ssu-rRNA: the first steps



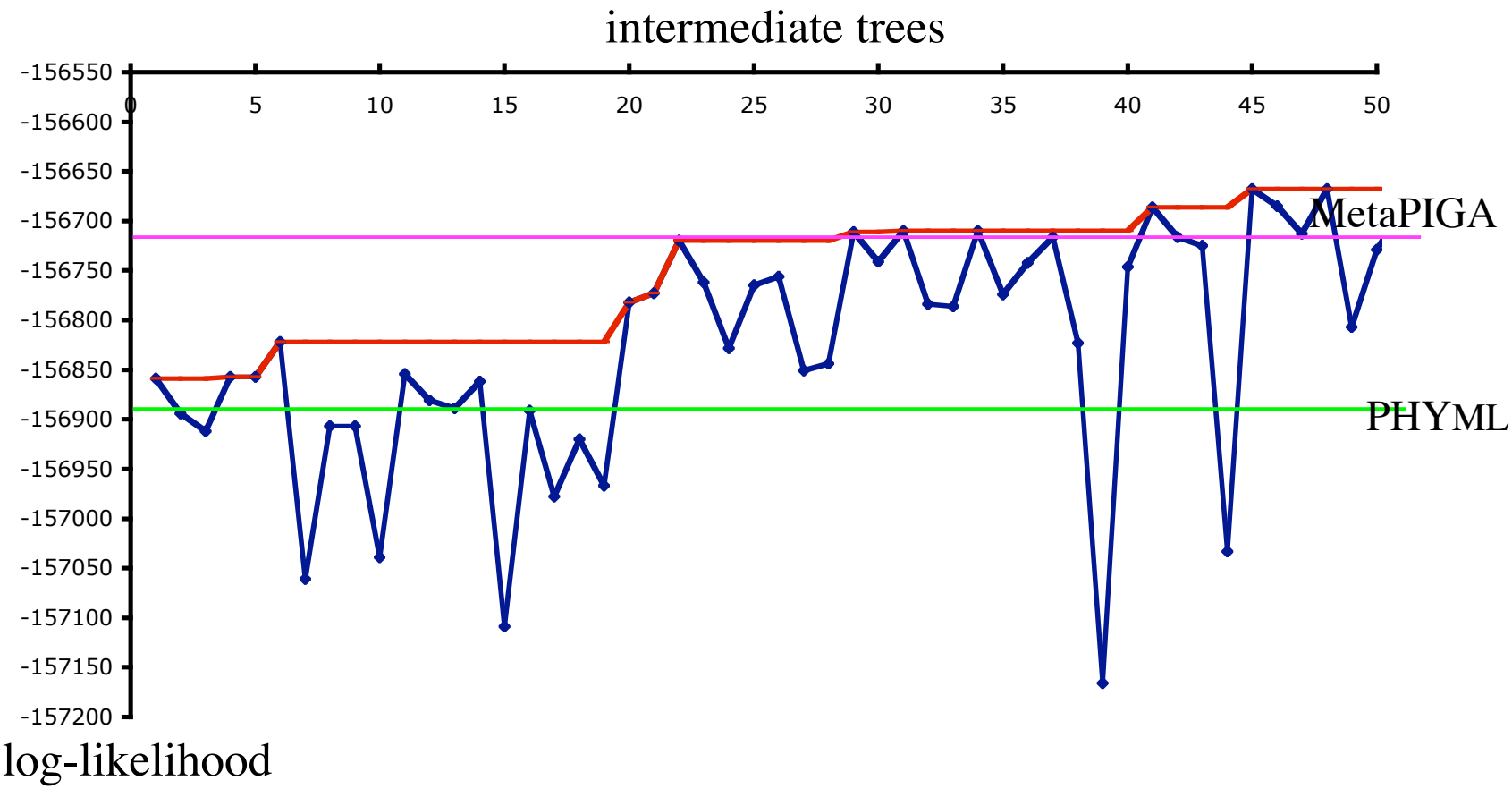
# ssu-rRNA: the first steps



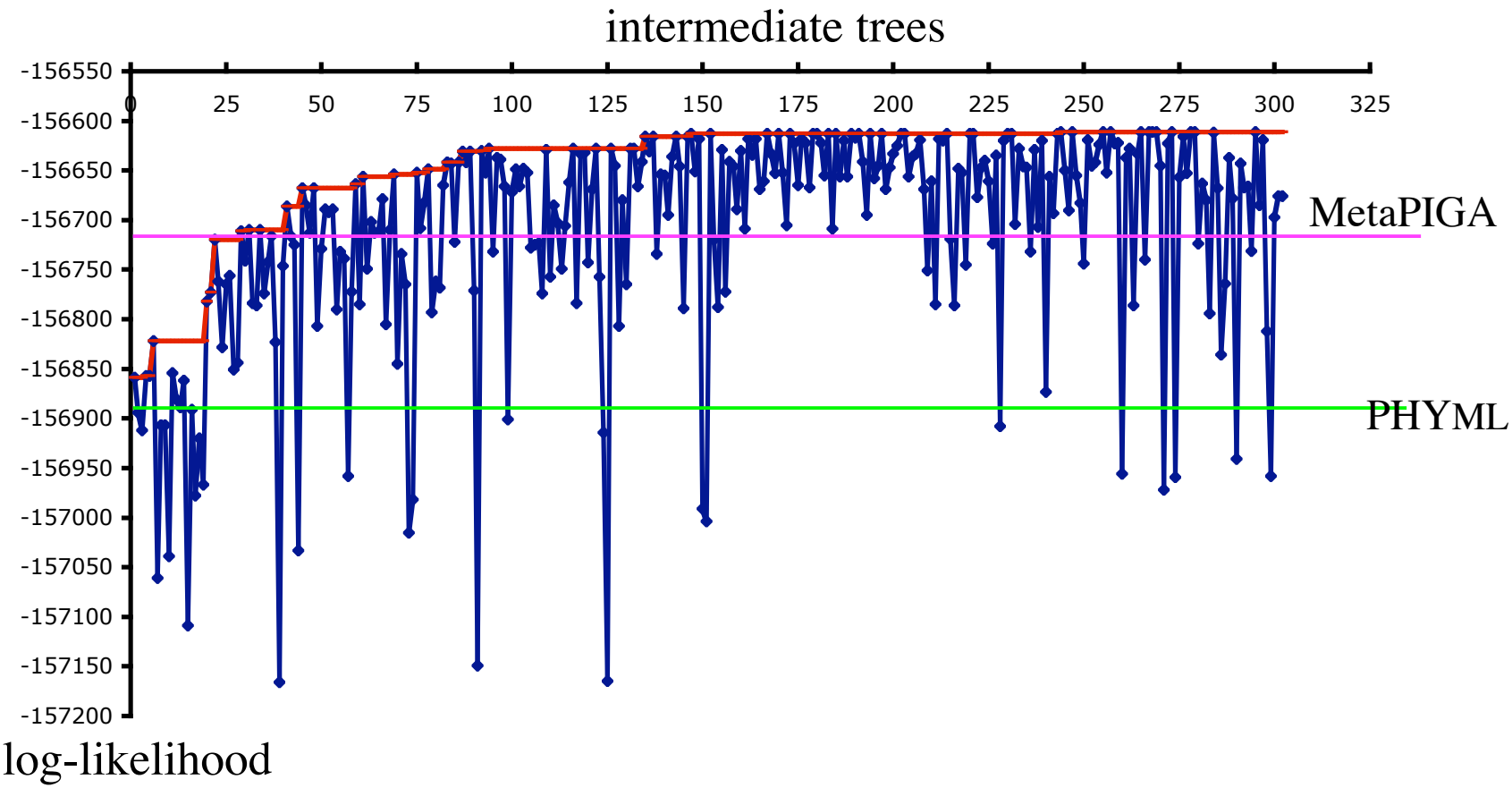
# ssu-rRNA: the first steps



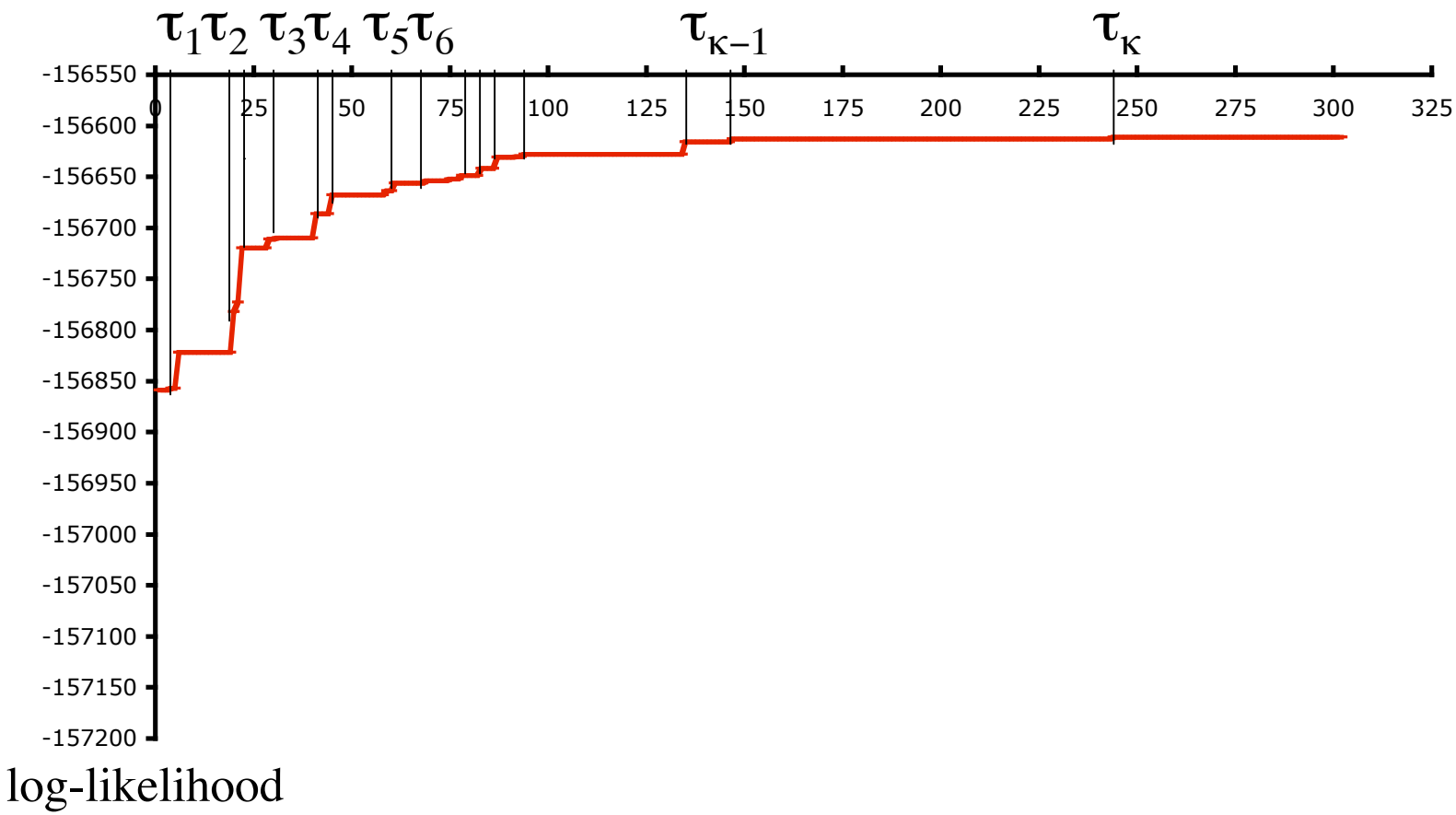
# ssu-rRNA: the quest continues



# ssu-rRNA: the final



# When should we stop?



# Help from the Dodo

**Figure 1** Dead as a dodo: the flightless bird from Mauritius and the adjacent islands weighed in at about 23 kg and was hunted to extinction. Its last confirmed sighting was in 1662, although an escaped slave claimed to have seen the bird as recently as 1674. In fact, it is estimated by using a Weibull distribution method that the dodo may have persisted until 1690, almost 30 years after its presumed extinction date. Although gone forever, the dodo's lumbering appearance in Lewis Carroll's *Alice's Adventures in Wonderland* has ensured that it will not be forgotten.

**When did the dodo become extinct?**

Robert DL, Solow AR (2003)
Nature 426:245-245

# Stopping time[1,2]

We use ($\tau_j$) to estimate the number of iterations to conclude with $(1-\alpha)100\%$ confidence that a further search will not be successful

1. Estimate shape parameter of a Weibull distribution

$$\hat{\nu} = \frac{1}{k-1}\sum_{j=1}^{k-2}\log\left(\frac{\tau_1 - \tau_k}{\tau_1 - \tau_{j+1}}\right)$$

2. Estimate of stopping time

$$\tau_{(1-\alpha)100\%} = \tau_1 + \frac{\tau_1 - \tau_k}{\left(\frac{-\log(\alpha)}{k}\right)^{-\hat{\nu}} - 1}$$

[1]Cooke P (1980) Biometrika 67:257-258
[2]Robert DL, Solow AR (2003) Nature 426:245-245

# What does it cost ?

| | | PHYML | MetaPIGA | IQPNNI | $\tau_{95\%}$ |
|---|---|---|---|---|---|
| gene | Number of sequences | Runtime (min) | | | |
| ssu rRNA | 218 | 5.1 | 74.5<br>1.2 h | 379.0<br>6.3 h | <br>8.4 h |
| rbcl | 500 | 7.5 | 158.5<br>2.6 h | 672.0<br>11.2 h | <br>15.0 h |

# Summary

At the expense of additional computing time we are able to reconstruct trees with a higher likelihood compared to other approaches tested.

We suggest an approach based on extreme value theory, when to stop the current search.

The analysis of many intermediate trees gives some insights which groups are well supported by the data and which not.

IQPNNI works for DNA and amino acid sequences

# Advertisment

IQPNNI can be downloaded from

www.bi.uni-duesseldorf.de/software/iqpnni