PhyQuart-A new algorithm to avoid systematic bias & phylogenetic incongruence

Are directed quartets the key for more reliable supertrees?

Patrick Kück

Department of Life Science, Vertebrates Division, The Natural History Museum London

Bioinformatics 2016

Tree Reliability & Long-Branch Attraction

Systematic errors in phylogenetics

- Increasingly apparent as more data are analysed
- Yielding maximally support of incorrect relationships
- Long-branch attraction (LBA) as a major source

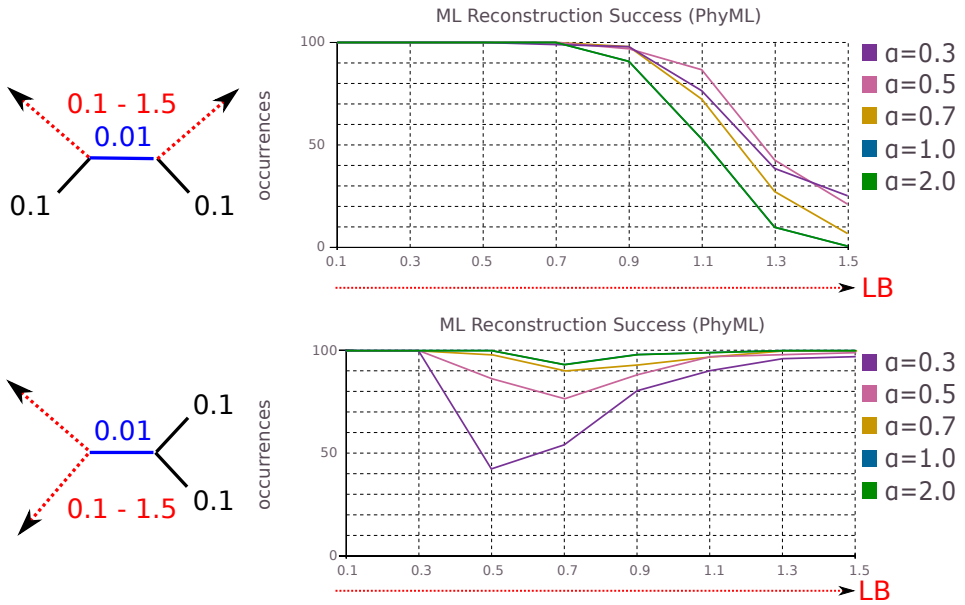
Which Topology is correct?

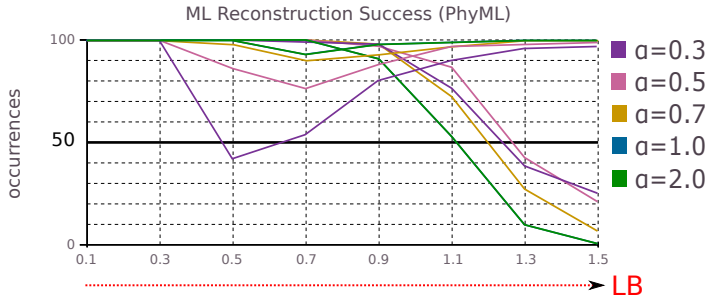
Terminal nodes can consist of...

- ...single taxa
- ...multiple taxa clades

Tree Reliability & Long-Branch Attraction

Maximum Likelihood Success (PhyML)

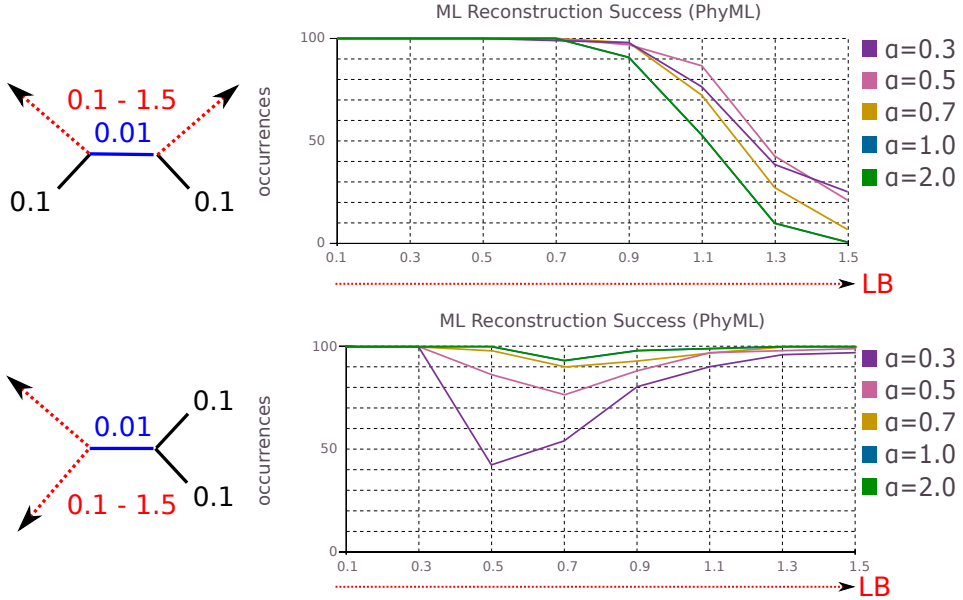


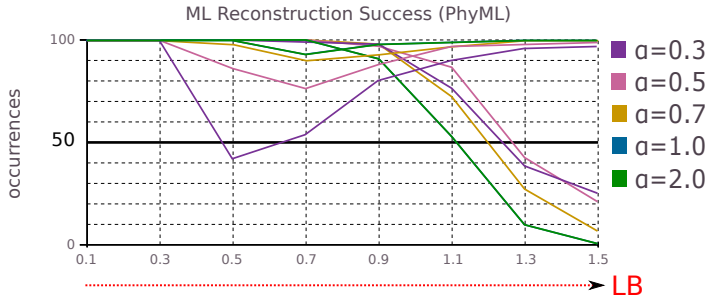


- GTR; α: 0.3, 0.5, 0.7, 1.0, 2.0; I: 0.3; L: 250.000bp
- 4 rate categories instead of continuous rate distribution for ML

Tree Reliability & Long-Branch Attraction

Maximum Likelihood Success (PhyML)





- ML Reliability further reduced by...
  - ...alignment errors
  - ...stochastic sampling errors
  - ...stronger model misspecifications

Tree Reliability & Long-Branch Attraction

Is it possible to develop alternative techniques that are less effected by extreme branch length asymmetries?

Tree Reliability & Long-Branch Attraction

Is it possible to develop alternative techniques that are less effected by extreme branch length asymmetries?

- Modern probabilistic substitution models assume time-reversibility
- Distinction between new (apomorphic) and old (plesiomorphic) homologies

NATURAL HISTORY MUSEUM

Tree Reliability & Long-Branch Attraction

Is it possible to develop alternative techniques that are less effected by extreme branch length asymmetries?

- Modern probabilistic substitution models assume time-reversibility
- Distinction between new (apomorphic) and old (plesiomorphic) homologies

PhyQuart

- Quartet based algorithm
- Consideration of 2 different directions of character alteration along the internal branch
- Allows discernibility between old and new character split-supporting site patterns and ...
- ... ML estimation of the expected number of convergent split support
- Combination of Hennigian logic and ML estimation represents a completely new strategy for the evaluation of sequence data

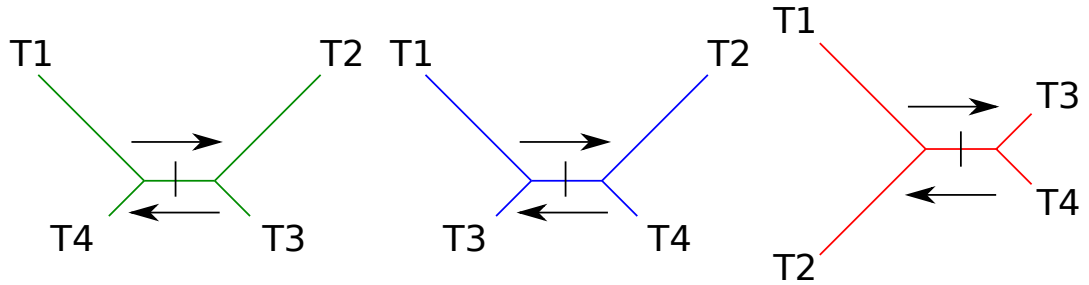
NATURAL HISTORY MUSEUM

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

3 Possible Quartet Trees for a Set of 4 Taxa

- 15 different split pattern

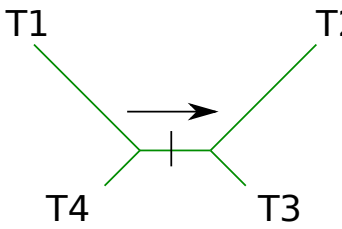
T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y
	Symmetric			Directive		Asymmetric						Singelton			



PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

3 Tree Supporting Split-Pattern

T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y
	Symmetric			Directive				Asymmetric				Singelton			



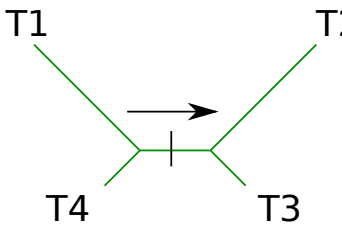
$N_{ap} = N_{tot}$ -

- N_{ap} : Potentially phylogenetic informative split-pattern signal
- N_{tot} : Total number of tree supporting split-pattern (alignment observed)

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

1 Uninformative, Old Split-Pattern per Tree Direction

T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y
	Symmetric			Directive				Asymmetric				Singelton			

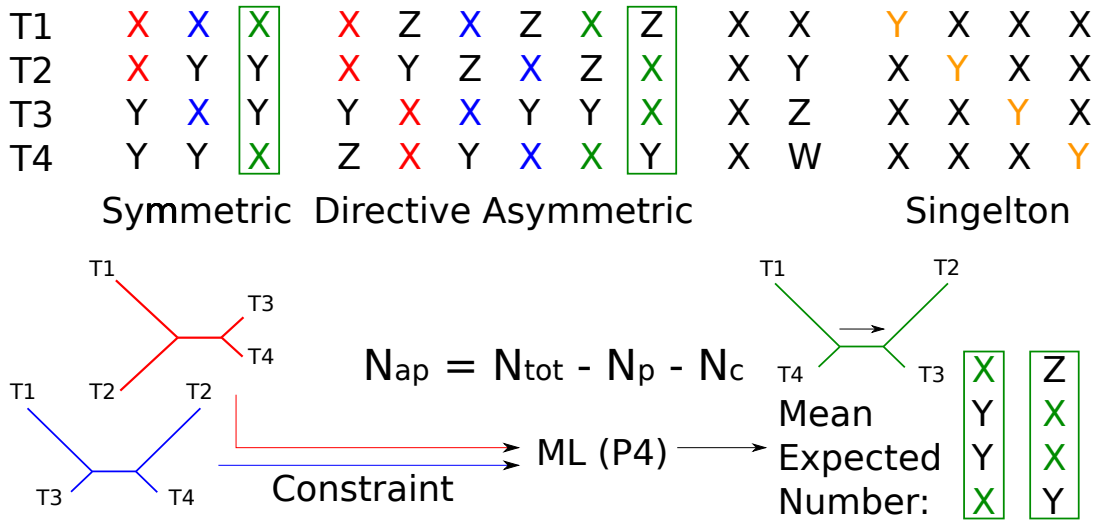


$N_{ap} = N_{tot} - N_p -$

- N_{ap} : Potentially phylogenetic informative split-pattern signal
- N_{tot} : Total number of tree supporting split-pattern (alignment observed)
- N_p : Plesiomorphic character similarity, uninformative (alignment observed)

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

2 Possibly Convergent Evolved Split-Pattern per Tree Direction

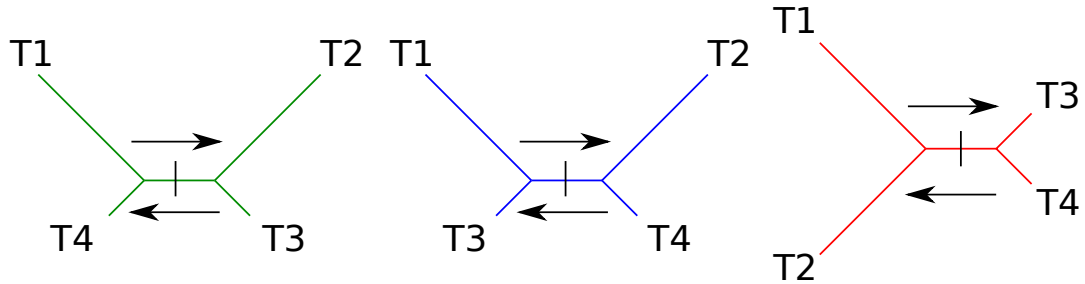


- N_{ap} : Potentially phylogenetic informative split-pattern signal
- N_{tot} : Total number of tree supporting split-pattern (alignment observed)
- N_p : Plesiomorphic character similarity, uninformative (alignment observed)
- N_c : Convergently evolved, uninformative (ML expected mean)

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Reduction of Support Underestimation

T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y
	Symmetric			Directive			Asymmetric					Singelton			

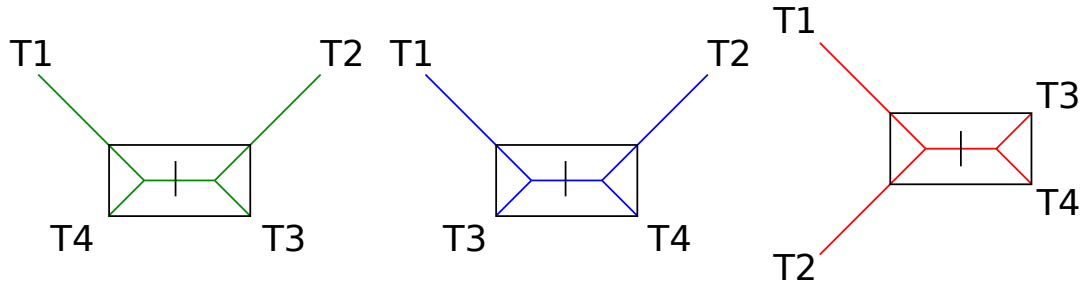


- Multiple hits may erode the support for the correct tree
- Correction of support values
- Frequency of singelton pattern as indicator for terminal branch lengths

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Reduction of Support Underestimation

T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y
	Symmetric			Directive			Asymmetric					Singelton			



Correction factor (CF):

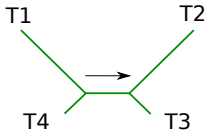
- $CF = (N_{Sing\_Smallest} * 4)/N_{Sing\_Total}$
- Corrected support values closer to what would be expected if external branches were of equal length

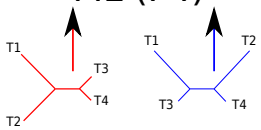
PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Reduction of Support Underestimation

	Symmetric			Directive				Asymmetric				Singelton			
T1	X	X	X	X	Z	X	Z	X	Z	X	X	Y	X	X	X
T2	X	Y	Y	X	Y	Z	X	Z	X	X	Y	X	Y	X	X
T3	Y	X	Y	Y	X	X	Y	Y	X	X	Z	X	X	Y	X
T4	Y	Y	X	Z	X	Y	X	X	Y	X	W	X	X	X	Y

$$\overrightarrow{N_{ap}} = \underbrace{CF_{obs} * (N_{tot} - \overrightarrow{N_p})}_{\text{Alignment}} - \underbrace{CF_{exp} * \overrightarrow{N_c})}_{\text{ML (P4)}}$$





Correction factor (CF):

- $CF = (N_{Sing\_Smallest} * 4)/N_{Sing\_Total}$
- Corrected support values closer to what would be expected if external branches were of equal length
- 2 correction factors: CF_{obs} (Alignment) & CF_{exp} (ML)

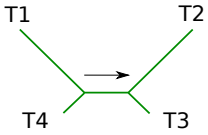
PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

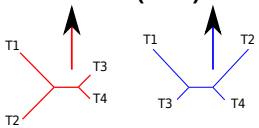
Final Scoring

PhyQuart Scores

$$\overleftarrow{N_{ap}} < \overrightarrow{N_{ap}} < \overrightarrow{N_{ap}} < \boxed{\overleftarrow{N_{ap}}} < \boxed{\overleftarrow{N_{ap}}} < \boxed{\overrightarrow{N_{ap}}}$$

$$\overrightarrow{N_{ap}} = \underbrace{CF_{obs} * (N_{tot} - \overrightarrow{N_p})}_{\text{Alignment}} - \underbrace{CF_{exp} * \overrightarrow{N_c})}_{\text{ML (P4)}}$$



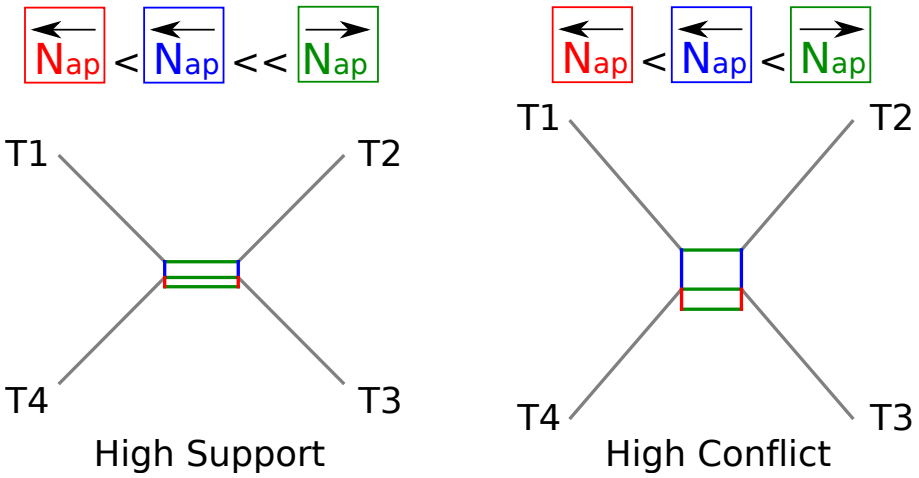


PhyQuart Score:

- For each quartet tree it's the highest of the scores for it's polarised quartets
- Normalised so that the scores of all three alternative trees sum to 1
- PhyQuart results imply both info about support scores & root info

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Final Scoring

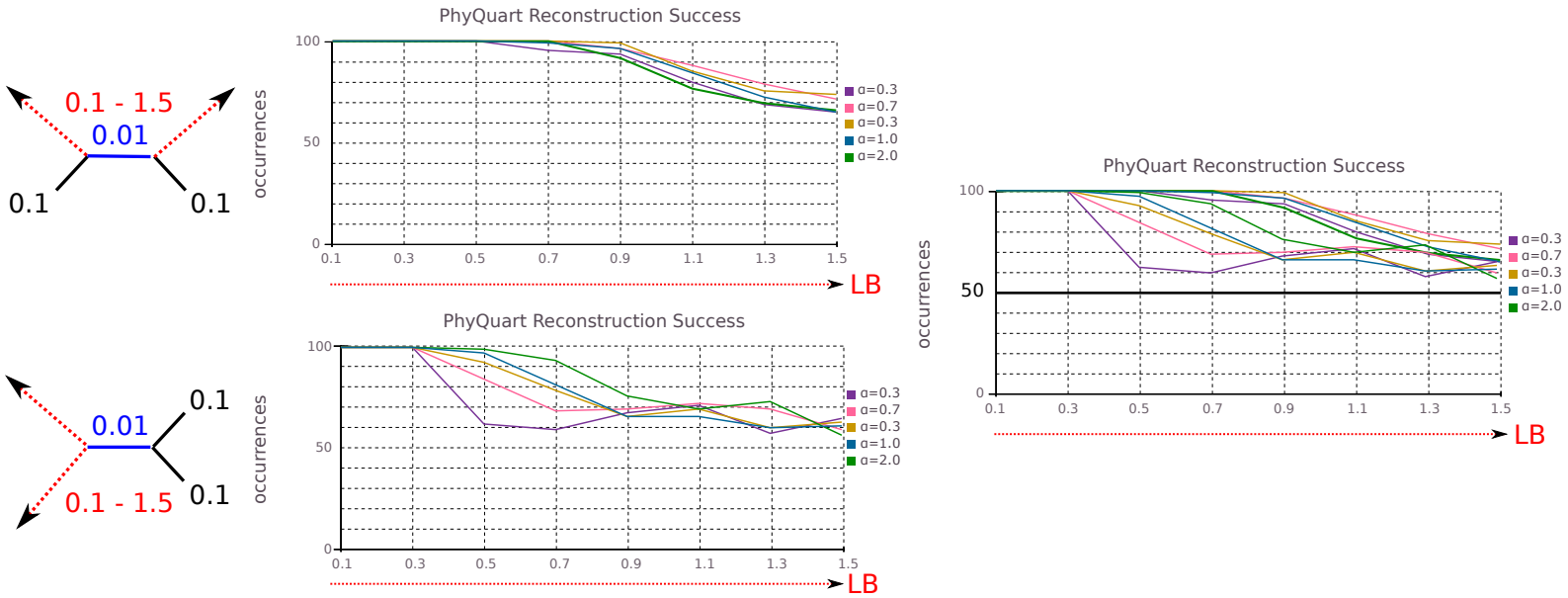


PhyQuart Score:

- For each quartet tree it's the highest of the scores for it's polarised quartets
- Normalised so that the scores of all three alternative trees sum to 1
- PhyQuart results imply both info about support scores & root info
- PhyQuart score network-graph

PhyQuart - Performance in Identifying Correct Quartets

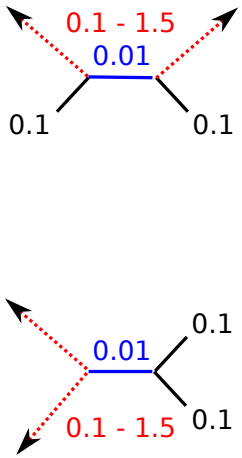
PhyQuart Success

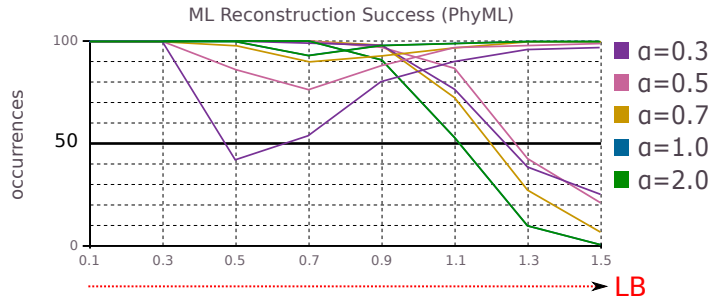


- GTR; α: 0.3, 0.5, 0.7, 1.0, 2.0; I: 0.3; L: 250.000bp
- 4 rate categories instead of continuous rate distribution for ML estimation

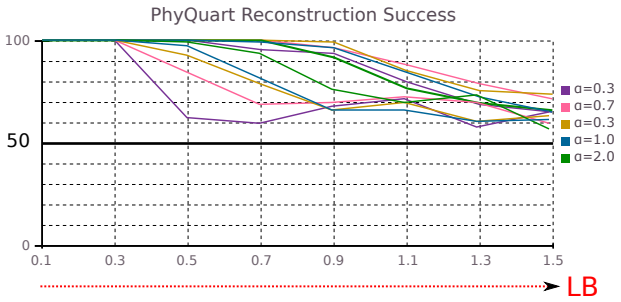
PhyQuart - Performance in Identifying Correct Quartets

PhyQuart Success





Maximum Likelihood



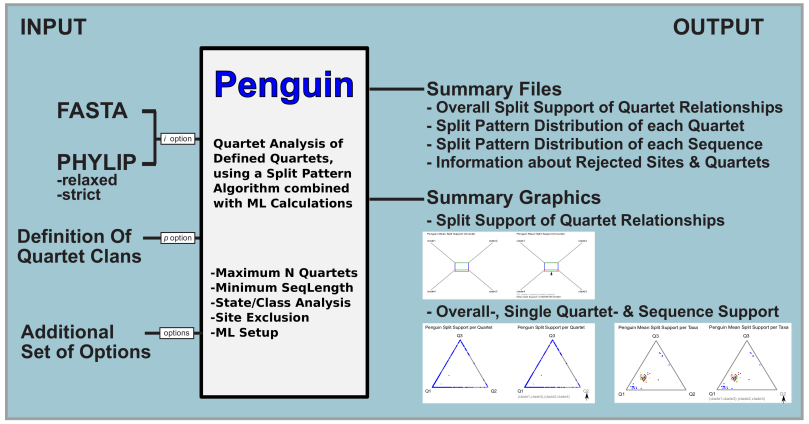
PhyQuart

PhyQuart . . .

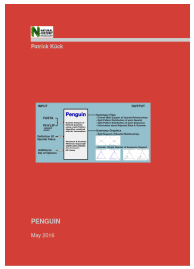
- . . . is quite successful in inferring correct quartet topologies from very heterogeneous sequence data
- . . . can outperform ML in both overcoming of long-branch attraction & repulsion
- . . . not recommended for shorter sequence lengths (<50 kbp)

Implementation of PhyQuart

PENGUIN



Manual

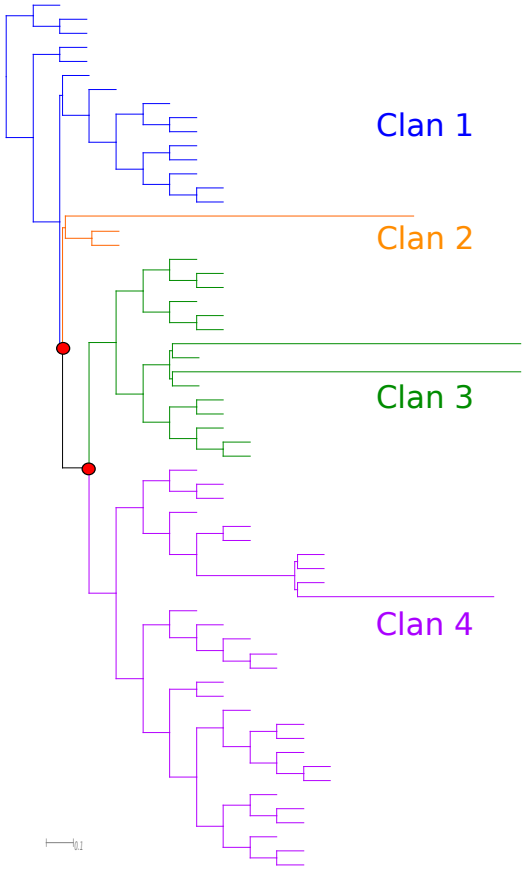


- Command line driven Perl script
- Runs on Windows, Mac OS, and Linux
- Extensive user options available
- Download Link: https://github.com/PatrickKueck/Penguin

NATURAL HISTORY MUSEUM

Applicability of PhyQuart (PENGUIN)

Divide & Conquer

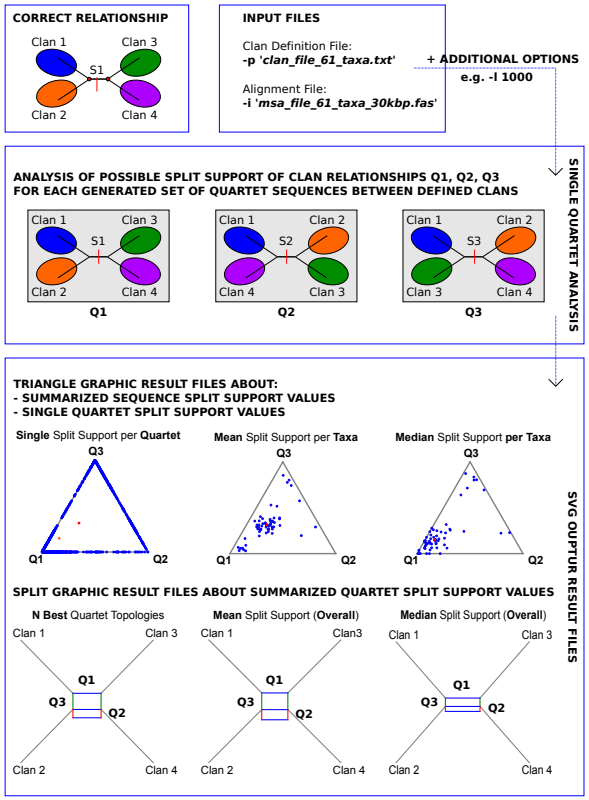


Analysis of...

- ...all quartets of larger trees
- ...predefined quartets of multitaxon clans

Applicability of PhyQuart (PENGUIN)

Divide & Conquer



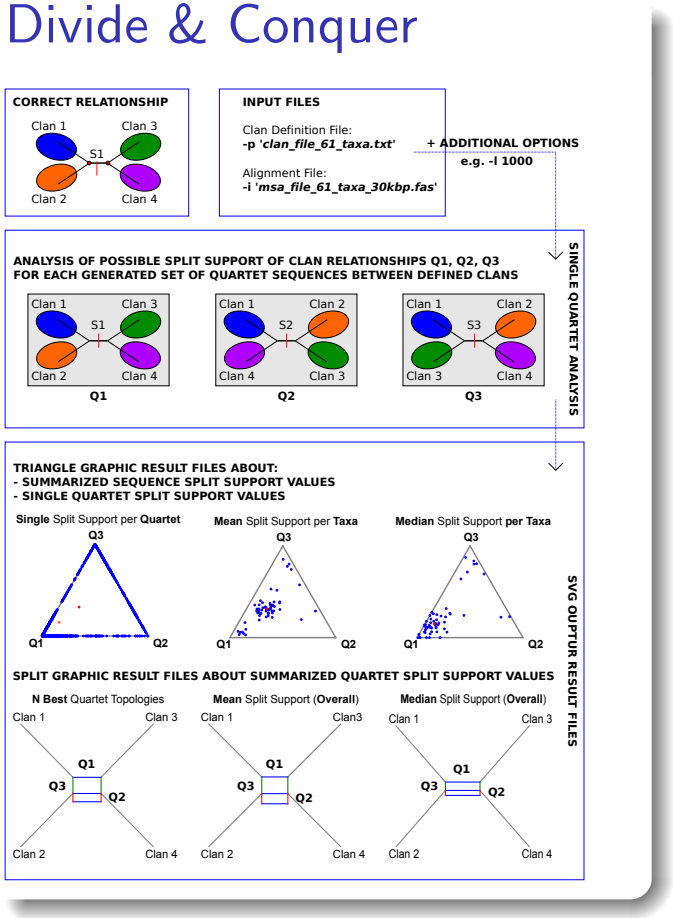
Analysis of...

- ...all quartets of larger trees
- ...predefined quartets of multitaxon clans

Evaluation of...

- ...contradicting signals to assess the robustness of relationships within a more complex tree

Applicability of PhyQuart (PENGUIN)



Analysis of...

- ...all quartets of larger trees
- ...predefined quartets of multitaxon clans

Evaluation of...

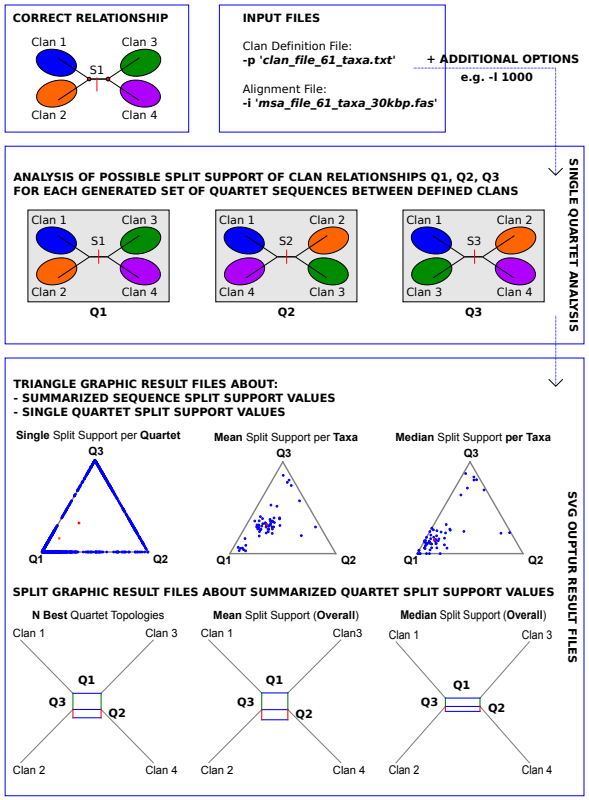
- ...contradicting signals to assess the robustness of relationships within a more complex tree

Identification of...

- ...of potentially rogue taxa

Applicability of PhyQuart (PENGUIN)

Divide & Conquer



Analysis of...

- ...all quartets of larger trees
- ...predefined quartets of multitaxon clans

Evaluation of...

- ...contradicting signals to assess the robustness of relationships within a more complex tree

Identification of...

- ...of potentially rogue taxa

Used...

- ...in combination with quartet-based supertree methods
- ...for network development

PhyQuart - Publication

Submitted to *Journal of Theoretical Biology*

Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data?

Patrick Kück*[1], Mark Wilkinson[1], Christian Groß[2], Peter G. Foster[1] and Johann Wolfgang Wägele[3]

[1] *The Natural History Museum, London, SW7 5BD, United Kingdom,* [2] *Pattern Recognition & Bioinformatics Group, Delft University of Technology, Delft, 2628 CD, The Netherlands,* [3] *Directorate, Zoologisches Forschungsmuseum Alexander Koenig, Bonn, 53113, Germany*

Thank you for your attention.