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
Introduction

Tree Reliability & Long-Branch Attraction

Maximum Likelihood Success (PhyML)

- GTR; $\alpha$: 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
Is it possible to develop alternative techniques that are less effected by extreme branch length asymmetries?
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 (apomorphous) and old (plesiomorphic) homologies
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 (apomorphous) and old (plesiomorphous) 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
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 | Z | Z | X | X | Y | X | X | X | X
T2 | X | Y | Y | Y | Z | X | Z | X | Y | X | Y | X | X
T3 | Y | X | Y | Y | X | Y | Y | X | Z | X | X | Y | X
T4 | Y | Y | X | Z | X | Y | X | X | Y | W | X | X | X | Y

Symmetric | Directive Asymmetric | Singelton
PhyQuart Algorithm

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

3 Tree Supporting Split-Pattern

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>T1</th>
<th></th>
<th></th>
<th>T2</th>
<th></th>
<th></th>
<th>T3</th>
<th></th>
<th></th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Z</td>
<td>X</td>
<td>Z</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>X</td>
<td>Z</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Z</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Z</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>W</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Symmetric  Directive  Asymmetric  Singelton

T1  T2  N_{ap} = N_{tot} -  T4  T3

- $N_{ap}$: 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

\[
\begin{array}{cccccccccccc}
T1 & X & X & X & X & Z & X & Z & \cancel{X} & Z & X & X & Y & X & X & X \\
T2 & X & Y & Y & X & Y & Z & X & Z & X & X & Y & X & Y & X & X \\
T3 & Y & X & Y & Y & X & X & Y & Y & X & X & Z & X & X & Y & X \\
T4 & Y & Y & X & Z & X & Y & X & \cancel{X} & Y & X & W & X & X & X & Y \\
\end{array}
\]

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

- \( T_1 \): X  X
- \( T_2 \): X  Y
- \( T_3 \): Y  X
- \( T_4 \): Y

**Symmetric Directive Asymmetric Singleton**

\[ N_{ap} = N_{tot} - N_p - N_c \]

- \( N_{ap} \): 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

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Z</td>
<td>Z</td>
</tr>
<tr>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Z</td>
<td>W</td>
</tr>
</tbody>
</table>

Symmetric  Directive Asymmetric  Singleton

Multiple hits may erode the support for the correct tree
Correction of support values
Frequency of singleton 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 | X | Y | X | X | X |
| T2 | X | Y | Y | X | Y | Z | X | Z | X | Y | X | Y | X |
| T3 | Y | X | Y | Y | X | X | Y | Y | X | Z | 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):

- \( CF = \left( N_{Sing_{Smallest}} \times 4 \right) / N_{Sing_{Total}} \)
- Corrected support values closer to what would be expected if external branches were of equal length
PhyQuart Algorithm

PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Reduction of Support Underestimation

Symmetric  Directive Asymmetric  Singelton

\[ \vec{N}_{ap} = \text{CF}_{\text{obs}} \times (\vec{N}_{\text{tot}} - \vec{N}_{p}) - \text{CF}_{\text{exp}} \times \vec{N}_{c} \]

Correction factor (CF):

- \[ CF = (N_{\text{Sing\_Smallest}} \times 4) / N_{\text{Sing\_Total}} \]
- Corrected support values closer to what would be expected if external branches were of equal length
- 2 correction factors: \text{CF}_{\text{obs}} (Alignment) & \text{CF}_{\text{exp}} (ML)
PhyQuart - Quartet Based Algorithm for Phylogenetic Inference

Final Scoring

PhyQuart Scores

\[ \text{Nap} = \text{CF}_{obs} \times (N_{tot} - N_p) - \text{CF}_{exp} \times N_c \]

Alignment

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; $\alpha$: 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

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**

**Command line driven Perl script**

**Runs on Windows, Mac OS, and Linux**

**Extensive user options available**

**Download Link:**
https://github.com/PatrickKueck/Penguin
Applicability of PhyQuart (PENGUIN)

Divide & Conquer

Analysis of...

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

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
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 Wilkinson¹, Christian Groß², Peter G. Foster¹ and Johann Wolfgang Wägele³

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

Thank you for your attention.