Announcements
1. Email me your list of 5 questions with correct answers before April 12th (last lecture) - I will post them on the website as a FAQs page
   - Must have date of question (approx. OK)
   - Question (proper spelling, grammar, etc.)
   - Answer (if answer is ambiguous, or unknown, do not use)
   - Worth bonus points of maximum 2%
2. Final paper deadline changed to April 12th

Troubleshooting Phylogenies
Outline
1. Parts of the tree seem “wrong”
2. Branch support values are low
3. Signals from different datasets or partitions conflict
4. Rooting the tree is problematic
5. Computation time is too long

Parts of the tree seem “wrong”
This speaks to the importance of knowing:
1. What all past investigations have suggested concerning the phylogeny
2. What morphology (& biogeography, etc.) suggest concerning the phylogeny
How else would one ascertain something “seems wrong”?
Probably many errors published because this information was not used

Example - my lab:
mtDNA (COI + COII)
Species seems to be in “wrong” place
Contamination?
Tube mix up?
Misidentification?
Copy-paste error?
How could one even suspect without morphology (or other information)?

Parts of the tree seem “wrong”
Either the results are wrong or the other information is wrong
- Always investigate both possibilities:
  - Problems with data / analysis (comparatively common)
  - Long-held traditional views on relationships, or morphological data, etc. are sometimes incorrect (comparatively rare)
Parts of the tree seem “wrong”

Onus is on the investigator to show results are strongly supported and alternative explanations have been rejected.

- Check for pre-processing errors:
  - alignment
  - use 2ndary structure or amino acid sequence
  - delete / ignore ambiguous regions
  - do analyses with different alignments of same data & compare
  - orthology / paralogy
    - use single-copy genes, gel-extract desired band
    - lab error (tube mix up, misidentification, copy/paste)
  - use multiple samples one level below level of interest

Parts of the tree seem “wrong”

2. Sample more genes with different substitution rates:
  - Compare faster with slower genes
  - deep level phylogenies - use slower genes
  - shallow level phylogenies - use faster genes
  - medium-level - use a mixture of genes of different rates
  - Compare nuclear with mitochondrial (or chloroplast)

Parts of the tree seem “wrong”

3. Check if rate heterogeneity is sufficient to cause LBA

“Detecting phylogenetic problems that are in the Felsenstein zone is a lot like detecting black holes;
such phenomena can only be inferred indirectly.”


Parts of the tree seem “wrong”

2. Are they joined in Parsimony analyses?

LBA due to model violation

With strong support …

But if MP doesn’t join - then LBA is rejected

1 of 4 shortest trees L=75
3. Are they joined by Maximum Likelihood methods?

- Model violation! - Make sure model is “best-fitting”

If not with ML but with MP
- then possible LBA affecting MP
- or relationship is correct (Farris zone)

If with ML & MP
- then possible LBA affecting MP & ML
- or relationship is correct

4. Do they join to different branches in each other’s absence?

Parsimony

If they join to the same place in each other’s absence it cannot be due to LBA

1 of 12 shortest MP trees

5. Are the branches long enough to attract?

This can be answered using simulated data

What if…

Data had evolved following these same branch lengths?

On a tree in which the long branches are separate?

Methods - example:

6 trees with the long branches separate
1 tree with them together
(built from 2 max LIL trees of MCMC with each long branch alone)

100 simulated DNA datasets per tree of lengths
50, 100, 250, 500, 1000 base-pairs

7 x 5 x 100 = 3500 per model, 3 models = 10500 analyses, took 170 CPU days

5. Are the branches long enough to attract?

Methods - example:

Data generated with program Seq. Gen. v1.27 using the simplest (and least realistic) model of evolution (Jukes Cantor)

equal base frequencies, no ASRV, equal substitution probabilities

Morphological data simulated from JC69 datasets - removed constant characters
Introduction to Biosystematics - Zool 575

Parts of the tree seem “wrong”
4. Assess the extent of Base Composition Bias & Heterogeneity - another source of phylogenetic error
- Some genes are biased in favor of AT or GC
- This compositional bias alone has explained some erroneous results
- Some cases are so bad even the use of amino acids is not free of the bias & yields erroneous results


Parts of the tree seem “wrong”
4. Assess the extent of Base Composition Bias & Heterogeneity


Correct tree:

ML tree - wrong due to base composition bias

Parts of the tree seem “wrong”
4. Assess the extent of Base Composition Bias & Heterogeneity

- Can compare base composition across taxa in various programs - standard to do so prior to any analysis
- Using PAUP* one can conduct a Chi-square test of homogeneity of base frequencies across taxa
- Can also compare a newer ML model that relaxes the assumption of homogeneity (stationarity) across taxa to one that assumes it (eg GTR) - use the Likelihood Ratio Test to compare (more on LRT in divergence dating lecture - see also ModelTest documentation & Purada Chapter)

PAUP* - test of homogeneity

Chi-square test of homogeneity of base frequencies across taxa:

<table>
<thead>
<tr>
<th>Taxon</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>ausungbun</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>davidat</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>euryost</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>giraudes</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>hertoix</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>luciant</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>monascus</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>mollisost</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>18</td>
</tr>
</tbody>
</table>

Chi-square = 128.21330 (df=27), P < 0.000000

Message: This test ignores correlation due to phylogenetic structure.
4. Assess the extent of Base Composition Bias & Heterogeneity

If the hypothesis of homogeneity is not rejected:
- the taxa (OTUs) are said to exhibit stationarity or base compositional equilibrium
- stationarity more likely among closely related taxa

5. Test for a Covariotide / Covarion Effect

- rates of variation may vary not only among sites and among lineages but among both simultaneously
- eg some sites are invariant in certain taxa but free to vary in others (ie the ASRV/ correction which assumes the same correction applies to all taxa may be wrong)
- Covariotide (DNA) covarion (protein) effect
- Lockhart et al. (1998) - inequality test for this

Branch support values are low

1. Be aware of known BS bias
- Compare to Bayesian posterior probabilities
- Use bootstrap correction factors (not really an option - too computationally complex - often requires squaring the number of replicates)

2. Gather more data
- More genes (characters)
- More taxa (big literature on comparison of benefits of more taxa versus more genes)

Troubleshooting Phylogenies

Outline
1. Parts of the tree seem “wrong”
2. Branch support values are low
3. Signals from different datasets or partitions conflict
4. Rooting the tree is problematic
5. Computation time is too long


Mitochondrial genome tree
All blank nodes = 100%
Branch support values are low

3. Check for Rogue Taxa
- One or more OTUs that are unstable can destabilize an entire tree or portions thereof
- Can be unstable due to missing data (eg fossils) or being long branches (high rates of change)
- In such cases the closest relative of the rogue taxon is ambiguous
- Try consensus agreement subtrees

Long branch rogue taxon - low PP values

Same analysis without the rogue taxon - PP values increase throughout tree

<table>
<thead>
<tr>
<th>Gene</th>
<th>length</th>
<th>MP</th>
<th>ML</th>
<th>NJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2s rRNA</td>
<td>1,111</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1,786</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ATPase6</td>
<td>887</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ATPase8</td>
<td>207</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>COI</td>
<td>1,560</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>COII</td>
<td>705</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND1</td>
<td>785</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>CYTB</td>
<td>1,149</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>NADH1</td>
<td>981</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NADH2</td>
<td>1,047</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NADH3</td>
<td>956</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NADH4</td>
<td>1,307</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>NADH4L</td>
<td>1,307</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>NADH5</td>
<td>1,666</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NADH6</td>
<td>561</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Identical to genome tree: 2/15 5/15 3/15

Branch support values are low

4. Avoid using “Fast” bootstrap methods
- PAUP* and other programs allow one to use shortcuts to speed up bootstrapping
- These involve doing less rigorous heuristic searches - typically using only a starting tree and no branch swapping
- Shown to lower BS values, especially in larger trees

5. Bootstrapping with ML & Bayesian - use correct models & experiment with model complexity

Troubleshooting Phylogenies

Outline
1. Parts of the tree seem "wrong"
2. Branch support values are low
3. Signals from different datasets or partitions conflict
4. Rooting the tree is problematic
5. Computation time is too long


### Signals from Different Data Sets or Partitions conflict

1. **Compare signals of different partitions**
   - Strong but contradictory signals in different data partitions can lead to low branch support & errors
   - Most commonly used method is the ILD (Incongruence Length Difference) test
   - Compares the difference in tree lengths between a tree based on both partitions with the sum of the tree lengths of each partition on its own tree:
     \[ D_{xy} = L_{t(x)} - (L_x + L_y) \]
   - \( D_{xy} \) = length difference

2. **Resolving conflicts among data partitions**
   - Conflict is frequently found as datasets grow larger
   - Some (cladists, eg Kluge 1989), argue to *always combine* regardless of signal conflict
   - The justification being that the parsimony tree is an explanation of all the data (and has thus been more rigorously tested for rejection - the least falsified hypothesis)
   - Unresolved tree from combined analysis more honestly represents the data (since we can’t know which partition is correct)

### Troubleshooting Phylogenies

**Outline**
1. Parts of the tree seem “wrong”
2. Branch support values are low
3. Signals from different datasets or partitions conflict
4. Rooting the tree is problematic
5. Computation time is too long

---

### Rooting the Tree is Problematic

- An often overlooked but critical component
- Without rooting, character evolution cannot be inferred
- Need rooted trees to infer monophyly (and thus classification)
- Unrooted trees are found using most software - the root branch is chosen afterwards
- Multiple outgroups can be added to an analysis to test ingroup monophyly - the test is obviously stronger if the outgroups are close to the ingroups
### Rooting the Tree is Problematic

For example:
- If you are studying "reptiles" and use fish as an outgroup the "reptiles" will ALWAYS be found to be monophyletic
- If you use birds as an outgroup instead the test is stronger - if one of your outgroups is resolved as a member of your ingroup then ingroup monophyly is rejected

Best practice, if possible:
- Choose multiple outgroups as close to the ingroup as possible

### Rooting the Tree is Problematic

2. What to do when no obvious outgroup is known
- When group is poorly known or at the base of the tree of life
- Try midpoint rooting - assigns the root to the midpoint of the longest path between two terminal taxa
- Desperate method, may not identify correct root
- Compare midpoint rooting to rooting with alternate outgroups

### Rooting the Tree is Problematic

1. Dealing with distantly related outgroups
- Sometimes there is no outgroup close to your ingroup
- If they are very distant they may be no better than a randomized sequence (ie all historical information is gone)
- This can lead to Long Branch Attraction in which the longest ingroup branch joins to the root
- Test different outgroups, try them alone & in combination
- Try midpoint rooting...

### Troubleshooting Phylogenies

Outline
1. Parts of the tree seem “wrong”
2. Branch support values are low
3. Signals from different datasets or partitions conflict
4. Rooting the tree is problematic
5. Computation time is too long

### Computation Time is Too Long

- Datasets are getting larger
  - whole genomes are becoming available
  - 500 + taxon matrices for fewer genes
- parametric (model) methods are computationally complex
- Confidence measures like bootstrapping increases the computational time 100-fold or more
- "Besides taking increasingly long coffee breaks, several strategies may help."

---

**Computation Time is Too Long**

1. **Use Fast(er) Algorithms**
   - See lecture on large dataset analyses
   - heuristics modified to avoid being trapped in local optima
   - newer algorithms / programs built to run faster
   - run risk that accuracy / rigor is sacrificed (general tradeoff between rigor and time)

2. **Reduce Complexity of Parametric Models**
   - Use model selection criteria to choose a model that is not unnecessarily complex
   - Use a quick method to infer a ‘good’ tree, eg MP, and estimate ML parameters & fix these for subsequent analysis (best to iteratively fix & search, fix again using new tree, continue until parameters become stable)

3. **Sequence More Genes for the Same Taxa**
   - Although adding more taxa can increase accuracy it usually slows computational time
   - Adding more characters for the same taxa often reduces the number of equally good trees & speeds up the searches (makes the data more decisive)
   - Only true if additional characters have same signal

4. **Estimate the “Consensus Support” tree rather than the optimal tree**
   - time otherwise spent on search for optimal tree is saved
   - search only for estimate of strength of signal in data
   - eg Bayesian Inference MCMC
     - Sanderson & Shaffer seem biased: such methods “probably do not provide good estimates of the optimal tree”
     - odd statement in light of the fact that Bayesian MCMC typically finds same topology as ML

5. **Use Faster Hardware**
   - high performance computing - chips speeds double every 1.5 years (Moore’s Law)
   - parallel processing is becoming more widely available
     - parallelized software for MP, ML, and BI MCMC is available
   - all CPU nodes should be of equal speed - search will go as slow as slowest node

**Terms** - from lecture & readings
- base composition bias
- chi-square test of homogeneity of base frequencies across taxa
- stationarity / base compositional equilibrium
- LogDet Distances
- covariotide / covarion effect
- rogue taxa
- ILD test
- midpoint rooting
**Study questions**

What are four things to investigate if parts of the tree seem "wrong"?

What is the logic, when investigating possible LBA, behind testing to see if long branches join to different places when analyzed alone?

If long branches are found to be sister taxa when analyzed using ML and a best-fitting model, what can one say about the possibility of LBA artifacts?

McC racken & Sorenson (2005) dealt with a problem of short internal branches - they compared "hard" versus "soft" polytomies. Describe these two types of polytomies - which seems a better explanation for these author’s results?

---

**Study questions**

The mitochondrial genome of vertebrates includes over 16,000 sites - based on Cummins et al (1995) approximately how many (minimum) of these sites are needed to obtain >95% confidence for all nodes of their 10-species vertebrate phylogeny?

How might one determine if two or more data partitions are congruent and what alternative actions might be taken (and why) if they are found to be incongruent?

What is the best practice for rooting a tree?