Systematics - Bio 615

The 10 Willis (The commandments as given by Willi Hennig after coming down from the Mountain)

1. Thou shalt not paraphyle.
2. Thou shalt not weight.
3. Thou shalt not publish unresolved nodes.
4. Thou shalt not disperse, when vicariating will suffice.
5. Thou shalt not polyphyle.
6. Thou shalt not clique, nor keep company with those who do.
7. Thou shalt not pheneticize (unless thou desirkest to show how blasphemous it is).
8. Thou shalt not cluster, nor covet those algorithms that cluster.
9. Thou shalt not homoplase (any more than thou needest).
10. Thou shalt not intuit (for by doing thus, thou may arrivest at the unexplained truth).

Outline
1. Mechanistic comparison with Parsimony
   - branch lengths & parameters

2. Performance comparison with Parsimony
   - Desirable attributes of a method
   - The Felsenstein and Farris zones

Statistical Phylogenetics
- Grew from mathematical, computer, evolutionary, numerical studies, not as much from systematics

1. Parsimony (Cavalli-Sforza & Edwards, 1963)
   tree with minimum changes preferred

2. Maximum Likelihood (Cavalli-Sforza & Edwards, 1964)
   tree that maximizes probability of the data is preferred
   Made available for DNA based phylogenetics by Felsenstein in the early 1980s

ML comparison with Parsimony (MP)

Parsimony would never prefer the correct, but longer tree on the left whereas ML would (more on this later) - Parsimony also ignores autapomorphies

2 convergence events 1 convergence event

Outline
1. Mechanistic comparison with Parsimony
   - branch lengths & parameters

2. Performance comparison with Parsimony
   - Desirable attributes of a method
   - The Felsenstein and Farris zones
## Desirable Attributes of a method

1. Consistency
   - a method is **consistent** if it converges on the correct tree as the data set becomes infinitely large (i.e., there is no **systematic error**)
   - **all methods are consistent if their assumptions are met (i.e., their model is not violated)**
   - Conversely, model misspecification can cause any method to be inconsistent

2. Efficiency (cont.)
   - Some methods are consistent but not very efficient (require TONS of data to work)
   - MP, when its assumptions are not violated, is far more efficient than ML

3. Robustness
   - A method is robust if it is relatively insensitive to violations of its assumptions (how easily does it become inconsistent?)

4. Computational speed
   - Clustering methods are very fast, optimality criterion methods are much slower, and ML is the slowest method known
   - Clustering methods (e.g., NJ)
     1. Do not guarantee an optimal solution
     2. Do not permit the comparison of alternative solutions

5. Discriminating ability (cont.)
   - Ability to discriminate among alternative hypotheses - increases with decreasing speed
   - Problem: ML has far better discriminating ability than NJ but is so computationally intense it cannot evaluate as much tree space (cannot compare as many alternative trees) as faster methods like MP
Desirable Attributes of a method

6. Versatility
   - what kinds of data can be analyzed?
   - MP (before 2001) had this as a huge advantage over ML
   - mixed analyses of morphology & DNA
   - behavioral data
   - can weight different DNA characters
   - now (since 2001) we can do all this with ML
     - morphology, behavior, etc
     - mixed dataset analyses - different models for different genes or types of data

A comment on data

When a single dataset has multiple components - eg morphology and DNA
these are called partitions of the data
unweighted MP treats all character state changes, whether morphology or DNA, the same

A comment on data - Partitions

ML methods, esp Bayesian, allow us to assign different models to different partitions
This greatly increases the fit of the meta-model (combination of all sub-models) to the data
New mixture-model methods (Bayesian) allow the data to tell the investigator how many distinct data partitions are present (more on this later)

Back to performance attributes...

Choosing a method requires balancing the importance of all these attributes
eg clustering is good to get starting trees or to explore the data quickly before doing a longer analysis
eg MP might be used to explore tree-space thoroughly and obtain starting trees and parameters to feed into a slow ML search

Back to performance attributes...

But the attribute most relevant to phylogenetic accuracy is determining if the assumptions of the method are met by the data - Model Fitting!
If not, then one risks model misspecification and potential inconsistency (or at least poor performance) of the method
If not, then all the other attributes matter little (except the method’s robustness to violation of assumptions!)

Outline

1. Mechanistic comparison with Parsimony
   - branch lengths & parameters
2. Performance comparison with Parsimony
   - Desirable attributes of a method
   - The Felsenstein and Farris zones
The Felsenstein & Farris Zones

Felsenstein (1978) demonstrated that MP would be inconsistent in a region of tree space now called the “Felsenstein Zone” - Long Branch Attraction (LBA)

Siddall (1998) dubbed the opposite region of tree space, where he hoped ML would be inconsistent, the “Farris Zone” - Long Branch Repulsion

Results of simulations

With Felsenstein zone trees

Long Branch Attraction

- Felsenstein (1978) used simulated DNA evolved on a simple model phylogeny of four taxa with a mixture of short and long branches
- Parsimony will give the wrong tree [misled by convergence]

<table>
<thead>
<tr>
<th>Model tree</th>
<th>Parsimony tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Rates or Branch lengths</td>
<td>p = q</td>
</tr>
<tr>
<td>p &gt;&gt; q</td>
<td></td>
</tr>
</tbody>
</table>

- With more data the certainty that parsimony will give the wrong tree increases - parsimony is statistically biased and inconsistent “positively misleading”
- Cladists claimed that Felsenstein’s results were unrealistic
- Few empirical examples known

Long Branch Repulsion? Or lack of LBA?

- Siddall (1998) claimed to have demonstrated that when long branches are sister taxa ML will fail to find the correct tree more often than random choice (worse than random). He suggested the long branches repelled each other and ML was failing the same way MP fails
- He called this region of treespace in which ML supposedly fails and MP never fails “the Farris Zone” in honor of Steve Farris, a great advocate of MP
- Swofford et al. (2001) demonstrated, in reply:
  - That, ML, will, as theory predicts prefer the correct tree, given enough data
  - ie it is not statistically inconsistent
  - Repulsion does not happen for ML - and neither does LBA
  - MP prefers the correct tree (with artifactually inflated branch support) due to LBA - ie its bias works in its favor

Farris Zone - even JC69 shows consistency with enough data

MP succeeds because it interprets homoplaspy as homology

Felsenstein zone Tree
Farris zone tree

There is no “Long Branch Repulsion”
ML not “positively misleading”

Failures ->
MP chooses the wrong tree with greater support as # of sites increases

Swofford et al (2001)
Unknown whether this is due to model violation or over-fitting or both

BRANCH LENGTHS!
Let’s compare apples to apples
Maximum Parsimony
The tree maximizing the probability of the observed data given the model - so… what’s the model?
(important so we know when it may be violated)

Maximum Likelihood
The tree maximizing the probability of the observed data given the model

So.. What is the Parsimony model?
Various attempts have been made to describe a parametric Parsimony model (Farris 1973, 1977, Felsenstein 1973, 1977)
It has not been easy…
The most promising is that of Goldman (1990)
Penny et al. (1994)
Tuffley & Steel (1997)
Under the hood of Parsimony...

These proofs provide parametric Maximum Likelihood process models that behave identically to unweighted parsimony.

Tangent - Process partitions & ML

Different data partitions - assign each a different ML process model
= increased realism
= more parameters
= fewer data for each...

Tangent - Process partitions & ML

e.g. Mitochondrial genome
2 partitions:
protein-coding and non protein-coding DNA
14 partitions:
13 protein coding genes & non protein coding DNA
27 partitions:
13 protein coding genes
2 rRNA genes
22 tRNA genes
16,500 partitions?

Process Partitions & Parsimony

All ML models make the parsimonious assumption that all characters share a common set of branch lengths
i.e. a common mechanism (CM)

Tuffley & Steel (1997) showed that Parsimony does not assume a common set of branch lengths...
~ Each character is its own data partition with its own set of branch lengths!
i.e. no common mechanism (NCM)

Maximum Parsimony

As typically used-
Can be understood as a nonparametric short-cut to a ‘strange’ parametric model

MP estimates a separate parameter for every branch by site combination \((2n-3) \times (\text{# of sites})\)
= # of parameters increases as you add data

“Statistically Pathological”

Estimated Parameters for 20 taxa

Number of Parameters

Number of Data
Problems with MP that ML solves

1. Statistical inconsistency & strong bias
2. Over-fitting of the data & (model violation / under fitting?)
3. Ignorance of branch length information during inference (ignoring sometimes useful data)
4. Correction for unobserved changes

Kolaczkowski & Thorton (2004)

Simulated data that violated the common mechanism assumption
- ML assumes all the characters share a common set of branch lengths (read “a common history of rates of change”)
- They generated data in which half of the characters had the exact opposite branch lengths as the other (heterotachy, heterotachous data - “different speeds”)

Kolaczkowski & Thorton (2004)

Results:
- ML and BMCMC did well with average homogenous and single partition homogenous but fails with heterotachous data until the internal branch reaches 0.28 sub/site
- MP did well with average homogenous but fails with heterotachous and single partition homogenous data until the internal branch reaches 0.23 sub/site (ie MP is inconsistent with heterotachous data if the length is less than this…)

Kolaczkowski & Thorton (2004) - My comments:
- Did not evaluate branch support - ML may have been appropriately cautious
- LBA trees are rare in real datasets - although there are many cases of possible LBA and in some regions of the tree of life, LBA is a fairly common problem (e.g. basal eukaryotes)
- This study takes the rarity of LBA to a new level by combining two equal and 100% opposite LBA patterns into one dataset!
Kolaczkowski & Thorton (2004) - My comments:

- New **mixture-model** Bayesian methods will not only help find different partitions in the data… (program BayesPhylogenies by Pagel & Meade)
- Will help identify partitions of the data that appear to share significantly different branch lengths
- allows us to relax the assumption of a common mechanism without throwing it out altogether (MP) eg instead of saying there are no shared branch lengths we can say there are 2 different sets of branch lengths - what fits the data best?

Kolaczkowski & Thorton (2004) - My comments:

- K&T balanced the heterotachy, what if it was unbalanced? Say 80% / 20% ?
- And finally… K & T did not address the issue of model - fitting
- Are there any datasets that are better fit by MP than a common-mechanism model?
- unpublished results indicate that there aren’t - there is little to no harm in assuming a shared set of branch lengths when this assumption is false

---

Performance across a full range of heterotachy

When the internal branch is = 0.15 ML succeeds if the heterotachy is not balanced and one partition has >80%

Performance across a full range of heterotachy

When the internal branch is = 0.15 or 0.05 MP fails regardless of the heterotachy (because ?)

---

**ML and MP - final comment on branch lengths**

- MP assumes no common set of branch lengths
- ML assumes a common set of branch lengths
- MP will do as well as ML when there is no common set of branch lengths or when all the branch lengths are equal (ultrametric)
- But MP will do less well than ML in between these extremes

---

**ML and MP - final comment on branch lengths**

- What if data are simulated using the MP model?
- We would expect, at least in this case, that MP would fit the data better than a ML model…
- not the case: MP doesn’t fit the data better and doesn’t perform better than ML (JC69) [P. Lewis pers. comm.]
Weaknesses of ML

- Can be inconsistent if we use poorly fitting models (model misspecification)
- Even our most complex models might not be complex enough
  - LBA can be demonstrated to happen even with GTR+I+G if the data are really complex
- Very computationally-intensive. Might not be possible to examine all key models (substitution matrices, tree topologies, etc.) - in which case a faster but less rigorous method might actually do better [Bayesian!]

Terms - from lecture & readings

6 desirable attributes of a method
- consistency
- efficiency
- robustness
- computational speed
- discriminating ability
- versatility
- mixed datasets
- partitions
- mixture-models
- Felsenstein Zone
- LBA
- Farris Zone
- LBR?
- positively misleading
- Parsimony model
- Tuffley & Steel (1997) [TS97]
- common mechanism
- no common mechanism
- statistically pathological
- mosaic evolution
- heterotachous data (heterotachy)
- Phylogenetic accuracy

Study questions

In the list of desirable attributes for a method some require trade-offs, eg a method can be computationally fast but not very discriminating - list examples of these trade-offs & different methods (NJ, ME, MP, ML) that demonstrate them.

What is the single most important thing one can do to maximize one’s chances of finding the true tree?

Why does MP fail in the Felsenstein zone? And why does it succeed so well in the Farris zone?

What is Long Branch Repulsion? Has it been verified as a real problem?

If increased realism and model complexity can be achieved by increasing the number of partitions in the data - why not go to the maximum possible and have each character a separate partition?

Study questions

ML and MP make different assumptions about branch lengths - what are these? When the assumptions are violated (in both directions) what is the result for MP & ML? (how robust are they to violation of this assumption?)

What is mosaic evolution (heterotachy) and how is this related to the assumptions of ML & MP?

How can we identify cases of mosaic evolution and what would be the best method to analyze such datasets?