Four steps - each should be explained in methods

1. **Character (data) selection** (not too fast, not too slow) "Why did you choose these data?"

2. **Alignment of Data** (hypotheses of primary homology) "How did you align your data?"

3. **Analysis selection** (choose the best model / method(s)) - data exploration "Why did you chose your analysis method?"

4. **Conduct analysis**

Polarity

**Hennigian terminology**

nouns

- **apomorph** - derived character state
- **autopomorph** - derived character state not shared
- **synapomorph** - shared derived character state
- **plesiomorph** - ancestral character state
- **synplesiomorph** - shared ancestral character state

adjectives

- **apomorphic, autopomorphic, etc.**

Final Review

Outline

1. Four steps
2. Hennigian Terminology
3. Saturation, correction, models of evolution, ASRV
4. Maximum Likelihood
5. Computation time is too long

Lack of a reproducible method resulted in three major approaches: (Explosion in 1960s)

1. **Phenetics** - similarity / distances only, not evolution, not phylogeny

2. **Cladistics** - phylogeny inferred using characters & parsimony

3. **Statistical Phylogenetics** - phylogeny inferred using corrected data & "best fitting model" [both distance & character data]
Introduction to Biosystematics - Zool 575

Saturation

Saturation graph
as time proceeds DNA distances also increase, to a point of saturation

Models of Sequence Evolution

For distance data:
- distances are the branch lengths
- correction to observed (D) to yield estimate of actual (d)
- simplest correction: Jukes and Cantor 1969 (JC69)
  \[ d = -3/4 \ln (1 - 4/3 D) \]
  eg D = 0.119, d = 0.130
  eg D = 0.00482, d = 0.00483

Models of Sequence Evolution

Some features many models deal with:
1. The tendency of one state to change to another (rate)
2. The frequency of the different states (base composition)
3. Variation in rates of change among characters (ASRV)
   - some sites change a lot, others change little or not at all

Maximum Likelihood - variable branch length

As the branch length increases the probability of a base changing increases (and the probability of it staying the same decreases - of course, they can’t both increase!)

For a branch 3 times longer than our initial branch:
\[ P^3 = \begin{bmatrix} 0.93 & 0.06 & 0.01 & 0.00 \\ 0.07 & 0.96 & 0.00 & 0.00 \\ 0.01 & 0.00 & 0.90 & 0.09 \\ 0.00 & 0.00 & 0.09 & 0.91 \end{bmatrix} \]

Models of Sequence Evolution

Among Site Rate Variation (ASRV)

- if present in data and not accounted for analysis can be misleading - inconsistent, erroneous
- can be accommodated by adding extra parameters to any model
- \( \text{plnvar} \) - proportion of invariable sites
  [ eg JC+I ]
- \( \text{gamma} \) - distribution of ASRV controlled by one parameter, alpha
  [ eg GTR+I+G ]

The shape of the gamma distribution for different values of alpha.

If alpha = infinity, rate is equal for all sites (no ASRV)
Low alpha (eg 0.5 and lower) = high ASRV
High alpha (eg > 2) = low ASRV
Models of Sequence Evolution
More complex models usually fit the data better

Optimality Criteria - Given 2+ trees

Maximum Parsimony
The tree hypothesizing the fewest number of character state changes is the best

Maximum Likelihood
The tree maximizing the probability of the observed data is best

Correction doesn’t always work...
- Huge dataset of 124,026 characters
- Data corrected with GTR and LogDet
- 100% branch support
- STILL WRONG!

Parsimony & Maximum Likelihood
Parsimony & Cladistics
- Strict cladists typically use only parsimony methods & justify this choice on philosophical grounds
eg it provides the “least falsified hypothesis”

- Parsimony has also been interpreted as a fast approximation to maximum likelihood

Cavalli-Sforza and Edwards (1967:555), stated that parsimony’s “…success is probably due to the closeness of the solution it gives to the projection of the ‘maximum likelihood’ tree and parsimony ‘certainly cannot be justified on the grounds that evolution proceeds according to some minimum principle…”.

Maximum Likelihood (ML)
Slow incorporation into systematics
1. Computationally complex - limited early uses to small molecular datasets (<12 OTUs)
2. Molecular only (until 2001)
3. Slow development of good software
4. Slow development of objective means to choose process models

Now we have more sophisticated software, faster computers, more realistic models, and means to choose among models

ML is becoming a widely used method
Joseph Felsenstein and David Swoford instrumental in developing software to use ML

Maximum Likelihood (ML)
A statistical model-fitting exercise

Which model (process + tree [topology+branch lengths]) is most likely to have generated the observed data?

Maximizing the probability of the data given a model \( L = Pr(D|H) \)

Select process model using AIC etc. select tree & model parameters using ML
### Maximum Likelihood (ML)

**Goal:** To estimate the probability that we would observe a particular dataset, given a phylogenetic tree and some notion of how the evolutionary process worked over time.

**Probability of:**
- `seq1 aacg`
- `seq2 accg`
- `seq3 aaca`
- `seq4 aatg`

### Maximum Likelihood (ML)

We then can rank & choose among alternate trees by selecting the tree which is most likely to have given rise to our data = an optimality criterion.

**Probability of:**
- `seq1 aacg`
- `seq2 accg`
- `seq3 aaca`
- `seq4 aatg`

### Maximum Likelihood

ML says nothing about the likelihood of the hypothesis being true it simply allows us to select among many hypotheses based on which makes the data most likely.

**eg. Loud sounds coming from your attic (data)**
- Hypothesis 1: gremlins are bowling upstairs
- Hypothesis 2: gremlins are sleeping upstairs

Hypothesis 1 makes the data more likely than hypothesis 2 - but itself is highly improbable!

### ML comparison with Parsimony (MP)

1. The most parsimonious topology is often, but not always, the same as the maximum likelihood topology.
2. Parsimony does not correct the data for unobserved changes and so parsimony branch lengths are typically underestimates of actual branch lengths.
   - Parsimony branch lengths = the estimated number of changes that occurred & are mapped onto the tree after it has been found (not used during tree searching - only total tree length is used to select the best tree)

### Maximum Likelihood - branch lengths

This is key... (1) ML estimates “corrected” branch lengths (corrected for unobserved changes)

And (2) uses these branch lengths to adjust the probabilities of change.

Branch lengths are estimated in units of expected number of changes per character.

The probability of change is a product of mutation rate ($\mu$) and time ($t$).

Branches can be long due to large $t$, or higher $\mu$, or both
- (hard to tease apart $t$ from $\mu$) if all branches have same $\mu$ then the data fit a molecular clock (ultrametric).
ML comparison with Parsimony (MP)

3. Because Parsimony ignores branch length information during searches it is
   - much faster than ML
   - unable to use this information to help it find the most probable tree (some strict cladists argue that they do not use parsimony to find the most probable tree, they claim there is no connection between minimal tree length and probability of being correct, “truth is unknowable”)
   - especially unable to use branch lengths (∼ rates of change) to detect areas of the tree that are likely to experience higher rates of homoplasy than other regions

Desirable Attributes of a method

1. Consistency
2. Efficiency
3. Robustness
4. Computational Speed
5. Discriminating ability
6. Versatility

- Some can be assessed using simulations
- Data are simulated and later analyzed using a method to determine its attributes
Desirable Attributes of a method

1. Consistency
   - inconsistency manifests as the preference for the wrong tree as the data become infinitely numerous

2. Efficiency
   - how quickly a method obtains the correct solution (how many data it needs to work)
   - tradeoff between consistency & efficiency

Desirable Attributes of a method

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

5. Discriminating ability
   - ability to discriminate among alternative hypotheses - increases with decreasing speed

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

Desirable Attributes of a method

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?)

Desirable Attributes of a method

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
   - may limit its applicability (but see future lectures on Bayesian methods)

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

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]

- 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 treesspace 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
  - if it is not statistically inconsistent
  - Repulsion does not happen for ML - and neither does LBA
  - MP prefers the correct tree (with artificially inflated branch support) due to LBA - if its bias works in its favor

Results of simulations
With Felsenstein zone trees

Farris Zone - even JC69 shows consistency with enough data
MP succeeds because it interprets homoplasy as homology

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

- 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(s) 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.]
**ML and MP - final comment on parameters**

- Some think of the parameters in a model as assumptions - eg the Ts/Tv ratio

- This can be specified a priori in both MP and ML which would be an assumption

- OR we can estimate it from the data! (ML can do this)

- the model allows the Ts/Tv ratio to differ from 1.0 only if the data indicate it does - it does not force it to differ

- an assumption is when the method is forced to use a certain value, the assumption is relaxed when the method is free to use any value the data indicate is best

**Weaknesses of ML**

- Can be inconsistent if we use poorly fitting models (model misspecification) - *like the Parsimony model!*

- 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
  - but new, better models are being developed

- 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

  - See Lecture on Bayesian methods - combines speed with ML models