Phylogeny Analysis

Measuring Random vs. Systematic Errors

- *random errors* can be broadly categorized as problems of insufficient sampling.
- Need to test whether our tree is likely to be quite different if we had a complete sample.
- We also must consider the possibility of *systematic errors* in the data / our assumptions.

Bootstrap Test: check internal agreement among multiple characters

Characters

Chimp	..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla	..AGCATAGGGGTCAGGGGAAAGGCT..
Human	..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque	..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan	..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel	..AGCGGACCGGTAAGGAGAAAGGAC..

Each character yields independent information about the correct phylogeny. While it’s not expected that all subsets of the characters would yield the exact same tree, we’d like to know if there are major disagreements within the set of characters. The Bootstrap Test addresses this, by *resampling* the set of characters.

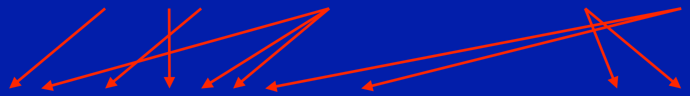
Bootstrap: make a random “sample” of characters with replacement

Characters

“original sample”

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Random sampling:



“new sample” is just random mix of old characters

```
Chimp      ..GACGGAACCCGGCAAAGAGGGGGG..
Gorilla    ..GTCGGTGCCCGGCAAAGAGGAGAG..
Human      ..GACGGAACCGGGGAAAGAGGGGGG..
Macaque    ..GCAGCCTAAAGGAAAAGAGGAGAG..
Orangutan  ..GCCGCCTCCAGGAAAAGAGGAGAG..
Squirrel   ..GGAGCGCAAAGGAAAAGAGGAGAG..
```

Sampling with replacement means each character can be sampled any number of times (0,1,2,3…).

Bootstrap: Sample Distance Matrix

Characters

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Distances

	Chimp	Gor	Hum	Mac	Orang	Sq
Chimp		5	2	8	8	9
Gor			5	7	6	8
Hum				9	8	9
Mac					2	4
Orang						5

Bootstrap: Sample Distance Matrix

Characters

```
Chimp      ..GACGGAACCCGGCAAAGAGGGGGG..
Gorilla    ..GTCGGTGCCCGGCAAAGAGGAGAG..
Human      ..GACGGAACCGGGGAAAGAGGGGGG..
Macaque    ..GCAGCCTAAAGGAAAAGAGGAGAG..
Orangutan  ..GCCGCCTCCAGGAAAAGAGGAGAG..
Squirrel   ..GGAGCGCAAAGGAAAAGAGGAGAG..
```

Distances

	Chimp	Gor	Hum	Mac	Orang	Sq
Chimp		5	2	11	8	11
Gor			7	9	6	9
Hum				11	8	11
Mac					3	3
Orang						6

Bootstrap samples yield a distribution of tree variants

- Features that are robustly preserved in trees produced from all bootstrap samples have high confidence.
- Compute Bootstrap Confidence: fraction of the bootstrap samples in which a given feature was preserved.
- If set of characters show divergent patterns, bootstrap will yield divergent results.

Monophyletic Group Confidence

- Use bootstrapping to generate a large number (e.g. 1000) of trees via resampling.
- To assess a putative monophyletic group (of sequences), ask what fraction of the trees contain it as a monophyletic group (i.e. a subtree containing no other sequences).
- Challenge: measure this fraction over all 1000 trees, for all possible monophyetic group permutations.

Consensus Monophyletic Group Extremes

- For a set of four sequences, what is the *maximum* number of consensus monophyletic groups possible?
- What is the *minimum* number of consensus monophyletic groups possible?

Consensus Monophyletic Group Extremes Answer

- The maximum is the case where every internal node in a rooted binary tree becomes a consensus monophyletic group. For four sequences, the maximum is **three**.
- The minimum is just the trivial case where only the entire set of sequences is (by definition) monophyletic. (i.e. the root of the tree is always present, so it always is a “consensus monophyletic group”).

Consensus Monophyletic Trees

- In a standard rooted tree, the nodes are monophyletic groups, and each monophyletic group (except the root) has an incoming edge from the smallest monophyletic group that contains it.
- We define a *consensus monophyletic group* as a monophyletic group with posterior probability above 50%.
- For example, we might identify them as the set of monophyletic groups that are found in more than half of a set of bootstrap samples for a specific set of sequences.
- We define a *consensus monophyletic tree* as a directed graph whose nodes are the consensus monophyletic groups for a given set of sequences, and each consensus monophyletic group (except the root) has an incoming edge from the smallest consensus monophyletic group that contains it. (Note that we can consider each leaf node (sequence) to be a consensus monophyletic group).

Monophy Algorithm

- Simple solution: monophyletic group reduces to *composition* (i.e. the set of children, irrespective of the subtree details)
- Represent composition as a *bitstring*, by assigning each sequence a unique bit.
- Combining two subtrees is just bitwise-OR.
- Traverse the tree with depth-first search, incrementing the count of each group found.

Monophy Algorithm

```
monophy(t, n): # analyze a tree or subtree
  if t is leaf: return bitcode[t]
  code1 = monophy(t->child1, n)
  code2 = monophy(t->child2, n)
  n[code1 OR code2]++
  return code1 OR code2
monophy_all(trees, n): # analyze all trees
  for tree in trees:
        monophy(tree, n)
```

Consensus Monophyletic Tree Extremes

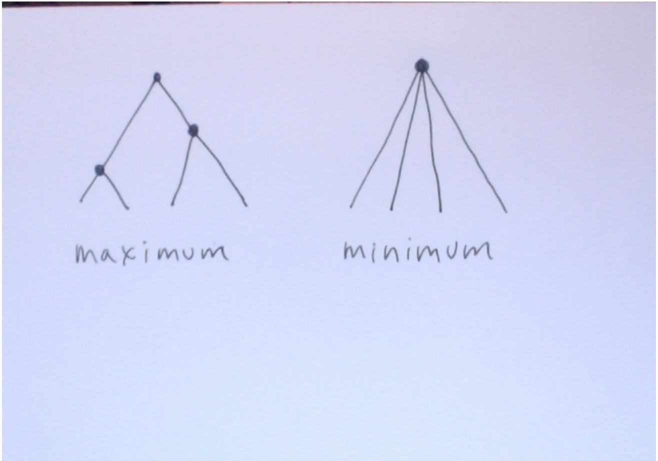
For a set of four sequences, draw a consensus monophyletic tree for each of these two scenarios:

- the consensus monophyletic tree with the *maximum* number of consensus monophyletic groups possible.
- the consensus monophyletic tree with the *minimum* number of consensus monophyletic groups possible.

Mark the locations of each consensus monophyletic group in the tree with a dot.

Consensus Monophyletic Tree Extremes Answer

- The maximum just looks like a standard binary rooted tree of 4 sequences.
- The minimum just has a single internal node (the root) with edges to all four of the sequences.

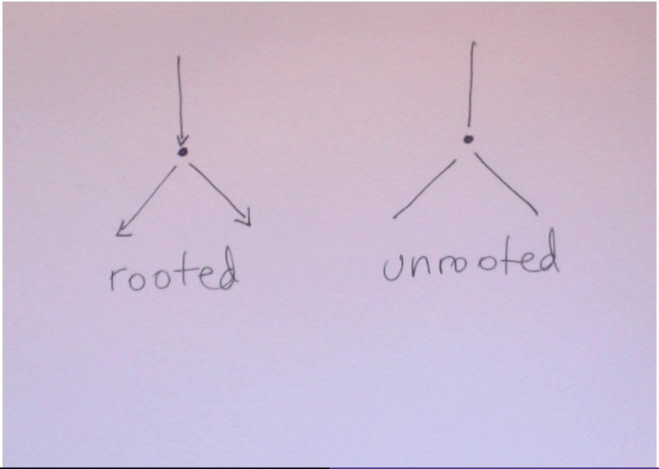


Unrooted Tree Subgrouping?

On a rooted tree we use an internal node to partition the tree into a monophyletic group vs. outgroup. Can this be done on an unrooted tree?

Unrooted Tree Subgrouping? Answer

No, we cannot use an internal node to partition an unrooted tree, because all three of its edges are equivalent (undirected), so we don’t know which edge to treat as “outgroup” (i.e. group the other two edges as a monophyletic group).



Unrooted Tree Edge “Cuts”

- An internal edge cuts an unrooted tree into two groups $G, \neg G$.
- We can refer to this by whichever group is smaller, e.g. G.
- An unrooted tree consists of a set of edges, where each edge is joined to the “smallest” group (edge) that contains it.
- We define a *consensus unrooted group* as an edge that has posterior probability $> 50\%$ (e.g. was found in more than half of the bootstrap samples).
- We define a *consensus unrooted tree* as an undirected graph whose edges are consensus unrooted groups, where each edge is joined to the “smallest” group (edge) that contains it. (We can consider external edges to be “consensus” by definition, since each one is present in *any* unrooted tree).

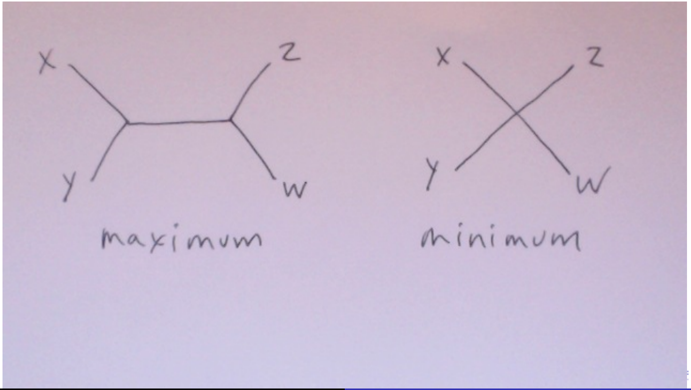
Consensus Unrooted Tree Extremes

For a set of four sequences, draw a consensus unrooted tree for each of these two scenarios:

- the consensus unrooted tree with the *maximum* number of consensus unrooted groups possible.
- the consensus unrooted tree with the *minimum* number of consensus unrooted groups possible.

Consensus Unrooted Tree Extremes Answer

- An unrooted tree for four sequences can only have one internal edge. If that edge had confidence > 50%, the consensus unrooted tree would just be the standard unrooted binary tree for four sequences.
- If that edge had confidence <50%, the consensus unrooted tree would have *no internal edges*.



Thanks!

- for taking on a new experiment in learning, and rising to the occasion!
- for all your hard work, pounding on all the questions in class, and all the work outside of class!
- for taking this in the spirit in which it was offered -- by giving your all.
- for teaching me more about teaching than my last 13 years at UCLA.
- for being the best class I've ever taught!