Stat 882: Statistical Phylogenetics – Lecture 1

Laura S. Kubatko
lkubatko@stat.osu.edu

Contents

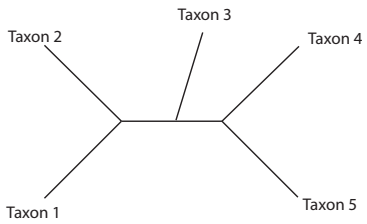
1 Introduction

1.1 Phylogenetic Trees

What is phylogenetics?

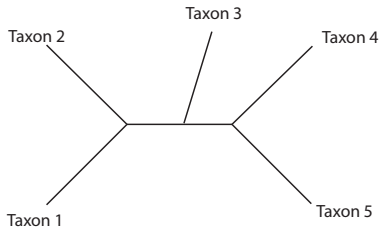
- Phylogenetics = the study of the evolutionary relationships among a collection of organisms – species, individuals, etc. – called taxa (singular taxon)

- We represent these relationships using a phylogenetic tree – a structure composed of nodes and branches
  - nodes = common ancestral organism
  - branches = ancestry-descent relationships



Intro to phylogenetics

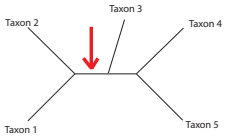
- External nodes = nodes at the tips of the tree. These generally represent present-day organisms.
- Internal nodes = represent ancestral organisms
- Often, branch lengths are taken as a measure of evolutionary time.

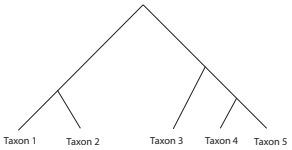


Intro to phylogenetics

- Trees are called rooted if the common ancestor to all taxa is identified.

- For example, we can add a root to our current tree to obtain a rooted tree.

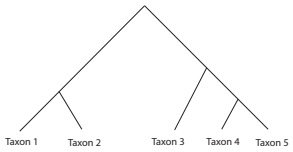


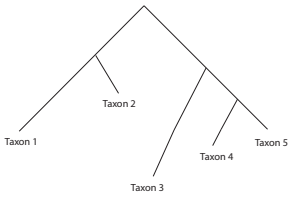


Intro to phylogenetics

- Rooted trees may or may not satisfy the molecular clock. The molecular clock assumption specifies that the time from each external node to the root is identical.

- Examples:



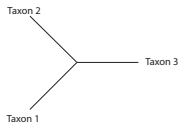


No molecular clock

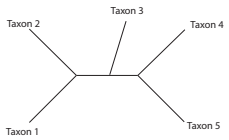
Intro to phylogenetics

- Generally consider only bifurcating trees (those with only three branches attached to each internal node)

- The number of trees grows rapidly in the number of tips – consider a tree with T taxa. There are
  - $T-2$ internal nodes
  - $2T-3$ branches
  - Number of possible trees is $\prod_{i=1}^{T-2}(2i-1)$



T=3, 1 internal node, 3branches, 1 possible tree

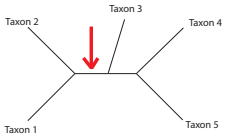


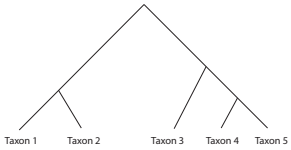
T=5, 3 internal nodes, 7 branches, 15 possible tree

Intro to phylogenetics

- For rooted trees, note that the addition of a root adds one more internal node and one more branch.

- The root can be placed along any of the previous $2T-3$ branches, so
  - the number of internal nodes increases by 1 to $T-1$
  - the number of branches increases by 1 to $2T-2$
  - the number of trees increases by a factor of $2T-3$ to $\prod_{i=1}^{T-1}(2i-1)$





Intro to phylogenetics

- Important point: The number of trees grows rapidly in the number of taxa:

Number of Taxa	Number of Rooted Trees	Number of Unrooted Trees
5	105	15
10	34,459,425	2,027,025
20	8.2×10^{21}	2.21×10^{20}
50	2.75×10^{76}	2.83×10^{74}

And recall our goal ..... to find an estimate of the phylogeny!

1.2 Data

Data used in phylogenetic analysis

- Two main types of data are used to construct phylogenies
  - Discrete character data
  - Distance (or similarity) data

- Discrete character data is probably most common. Sometimes, this type of data is transformed into distance data in order to estimate the phylogeny.

- Of course, continuous data are also used in evolutionary analyses involving closely related taxa. Here, however, the goal is often not to obtain a phylogenetic estimate but to carry out an analysis after adjusting for phylogenetic relationships. We'll discuss this at the end of the course.

Data used in phylogenetic analysis

- Character data: assume that we have a data matrix $\mathbf{X}$ of the form

	character 1	character 2	character 3	$\cdots$	character N
Taxon 1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1N}
Taxon 2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2N}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Taxon T	x_{T1}	x_{T2}	x_{T3}	$\cdots$	x_{TN}

where x_{ij} = character state for character j in taxon i

Data used in phylogenetic analysis

- General assumptions about discrete character data are:
  - Characters evolve independently
  - Characters are homologous: all states observed over taxa for that character are assumed to have derived from a corresponding state in the common ancestor of those taxa –May require alignment of characters (we'll talk about this process later)
- For now, we will assume these conditions are met, though we'll occasionally discuss relaxing the first condition

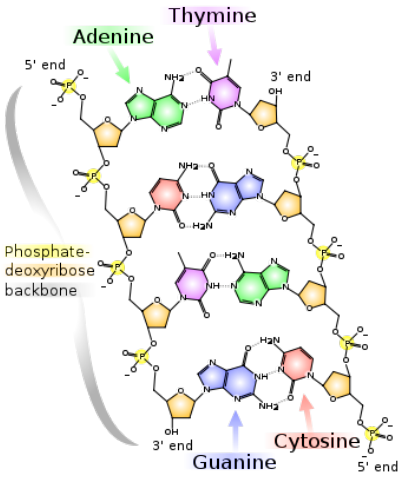
Data used in phylogenetic analysis

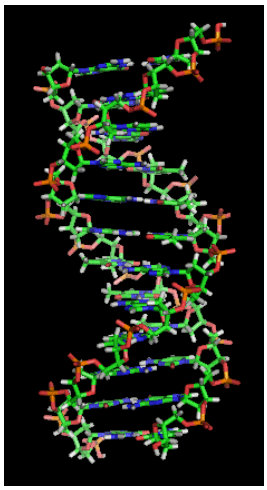
- Examples of data commonly used in phylogenetic analyses:
  - DNA sequence data
  - Morphological data
  - Allelic data
  - Gene order data

Example 1: DNA sequence data

- Biology background: DNA
  - DNA = deoxyribonucleic acid
  - molecule that contains the genetic instructions for the development and functioning of all living things
  - consists of strings of four possible nucleotides:
    - Adenine – A
    - Guanine – G
    - Cytosine – C
    - Thymine – T
    - A and G are called purines; C and T are called pyrimidines
  - backbone of sugar and phosphates holds the string of nucleotides together

Example 1: DNA sequence data





Data used in phylogenetic analysis

- Characters are positions in a set of aligned DNA sequences
- Character states are one of 4 possible nucleotides: A, C, G, or T
- Example data matrix (Rokas et al., 2003):

Taxon	DNA Sequence

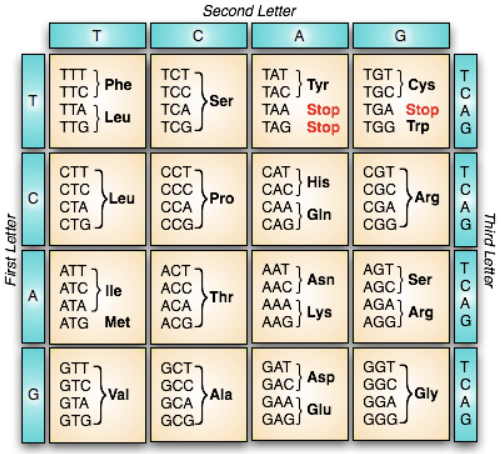
S. cerevisiae	T C	T	T T A T T G A C G T G T
S. paradoxus	T C	T	T T G T T A A C G T G C
S. mikatae	T C	C	T T G C T A A C A T G C
S. kudriavzevii	T C	T	T T G C T A A C G T G C
S. bayanus	T C	T	T T A C T A A C G T G C
S. castellii	T C	A	C T A T T A A C A T G T
S. kluyveri	T C	T	C T T C T A A C G T G C
C. albicans	T C	T	C T T T T G A C A T G T

Data used in phylogenetic analysis

- Examine assumptions in this case:
  - Assumption 1: Independence
    - Likely to be violated in coding regions of the DNA sequence. Sets of three nucleotides together make a codon, which codes for a particular amino acid (of 20 possibilities). These amino acids are strung together to produce proteins. So there are lots of constraints within coding regions.
    - Non-coding regions are less well-understood. Evolution of nucloetides is probably closer to independent across sites here than in coding regions.

Data used in phylogenetic analysis

- The genetic code



Data used in phylogenetic analysis

- Examine assumptions in this case:
  - Assumption 2: Sites are homologous
    - A crucial first step is alignment of the DNA sequences so as to attempt to meet this assumption
    - More on this later in the course
- Note that it is also reasonable to use amino acid sequences as the data in a phylogenetic analysis

Example 2: Morphological data

- Physical features of the organisms under consideration
- Often presence/absence of some feature of interest
- Could also classify features into categories
- Challenges:
  - How to choose characters?
  - Assumption of independence?
  - Assumption of homology across taxa?

Other Examples

- Allelic data:
  - Presence/absence of a particular version of a gene
  - Allele frequency data
- Gene order data: Arrangements of genes within genomes of individual taxa
  - Potentially very informative
  - Computationally very difficult

We will focus primarily on DNA sequence data in this course.

Phylogenetic Inference

- Goal: Given a data matrix $\mathbf{X}$, obtain an estimate of the phylogenetic tree. How?
- Algorithmic methods: define a sequence of steps that result in a bifurcating tree Example: Cluster pairwise distances until a tree has been built.
- Criterion-based methods: define an optimality criterion for comparing trees, and search for the tree that is optimal under the criterion.
  - Parsimony
  - Maximum likelihood
  - Some distance-based methods
- Bayesian methods

2 Parsimony

2.1 Parsimony: General Ideas

Intro to Parsimony

- Parsimony is one of the oldest and most common methods for inferring phylogenies
- Introduced by Edwards and Cavalli-Sforza in 1964 (see Ch. 10 in text for a nice account of the history of the field)

- Main idea of parsimony - simpler hypotheses should be preferred over more complex ones

- Thus, the tree required the fewest changes in character state should be preferred over other trees

- This tree is called the MP tree (most parsimonious or maximum parsimony)

Intro to Parsimony

- Advantages
  - Very general; can be applied to any data set for which we can quantify evolutionary change
  - Easy and intuitive to understand
  - In many situations, provides a good fit to the evolutionary scenario

- Disadvantages
  - Appropriateness of the assumption of fewest changes being most plausible?
  - Need a mechanism to handle ties between trees
  - Some others we'll see soon ....

The Parsimony Score

- Recall our previous notation: x_{ij} = character state for character j at node i
- Define $C(x_{ij}, x_{kj})$ to be the cost of changing from the state for character j at node i to the state for character j at node k over the branch connecting nodes i and k
- Note that we do not require $C(x_{ij}, x_{kj}) = C(x_{kj}, x_{ij})$, but this is most common
- Let nodes $1, 2, \cdots, T$ be the external nodes
- Let the internal nodes be denoted by $T+1, T+2, \cdots, 2T-2$
- Let N be the number of characters in the data set

The Parsimony Score

- With this notation, we define the parsimony score of tree τ to be

$$S(\tau) = \sum_{h=1}^{N} \sum_{b=1}^{B} C(x_{b_1,h}, x_{b_2,h}) \tag{1}$$

where B is the number of branches in the tree, and b_1 and b_2 are the nodes at the ends of branch b.

The Parsimony Score

- With this notation, we define the parsimony score of tree τ to be

$$S(\tau) = \sum_{h=1}^{N} \sum_{b=1}^{B} C(x_{b_1,h}, x_{b_2,h})$$

where B is the number of branches in the tree, and b_1 and b_2 are the nodes at the ends of branch b. Sum across branches: assumes independent evolution across branches, given states at the tips

The Parsimony Score

- With this notation, we define the parsimony score of tree τ to be

$$S(\tau) = \sum_{h=1}^{N} \sum_{b=1}^{B} C(x_{b_1,h}, x_{b_2,h})$$

where B is the number of branches in the tree, and b_1 and b_2 are the nodes at the ends of branch b. Sum across sites: assumes independent evolution across sites

The Parsimony Score

$$S(\tau) = \sum_{h=1}^{N} \sum_{b=1}^{B} C(x_{b_1,h}, x_{b_2,h})$$

- Can add a weight w_h in front of the cost to allow differential weighting of sites.
- Need to select cost function – a common choice is

$$C(x_{b_1,h}, x_{b_2,h}) = \begin{cases} 1, & x_{b_1,h} \neq x_{b_2,h} \\ 0, & otherwise \end{cases}$$

- This is called Fitch parsimony, and it can be applied to unordered multi-state data (DNA sequences, amino acid sequences, morphological data).

The Parsimony Score

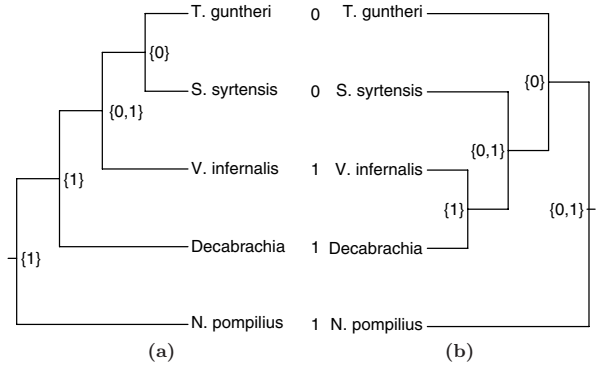
- Some other choices for the cost function:
  - DNA sequence data: assign lower cost to transitions than transversions
  - Codon data: assign lower cost to synonymous vs. nonsynonymous changes
  - multistate ordered character data: Wagner parsimony A change from any state to any other state incurs a cost that is equal to the sum of the intervening states.

The Parsimony Score: Example Data

- Data from Lindgren et al. (2004)
- Cephalopods – squids, cuttlefishes, octopi, etc.
- Class is divided into two groups: Nautiloidea and Coleoidea
- Coleoidea contains three subgroups:
  - Decabrachia – squids and cuttlefishes
  - Octobrachia – octopi
  - Vampyromorpha – vampire squid
- Placement of Vampyromorpha is controversial
- Lindgren et al. (2004) construct phylogenies from both morphological and sequence data

2.2 The Fitch Algorithm

Computing the parsimony score – The Fitch Algorithm



(a) (b)

Computing the parsimony score – The Fitch Algorithm

- Formalize the Fitch Algorithm
  - Initialize the external nodes of the tree by assigning a set containing the observed state
  - Visit nodes in the tree in a post-order traversal. At each node, consider the intersection of the sets at the two descendant nodes: If the intersection is non-empty, assign the intersection to the node under consideration. If the intersection is empty, assign the union of the two sets at the descendant nodes to the current node. Add one to the length of the tree.

Computing the parsimony score – The Fitch Algorithm

- Since the algorithm is designed for unordered multistate data, it doesn't matter which direction we move through the tree, which means that the tree can be arbitrarily rooted without affecting the length.

- For the example data, of 9 informative characters (characters whose length differs between the trees in (a) and (b)), 8 favor the tree in (a) and one favors the tree in (b).
- Under this criterion, the tree in (a) is favored.

2.3 The Sankoff Algorithm

Computing the parsimony score – The Sankoff Algorithm

- A second algorithm for computing the score of a tree under parsimony is the Sankoff algorithm.
- This algorithm works by assigning a function to each node of the tree which records, for each possible state, the minimum score for the subtree rooted by that node.
- Denote this function by $S_i^h(x)$, and define it to be the minimum score for the subtree rooted by node i assuming that node i has state x for character h.

Computing the parsimony score – The Sankoff Algorithm

- This value can be computed for any node for which this function has already been computed for its two immediate descendants, using the following relationship

$$S_i^h(x) = \min_{x_{j,h}}\{C(x_{i,h}, x_{j,h}) + S_j^h(x_{j,h})\} + \min_{x_{k,h}}\{C(x_{i,h}, x_{k,h}) + S_k^h(x_{k,h})\}$$

where j and k are the two nodes directly descending from node i.

The computational complexity is $O(n^2)$, where n is the number of possible character states (because finding the min is $O(n)$ and this is carried out n times at each node). The algorithm can be made more efficient – see Clemente et al., BMC Bioinformatics 2009, 10:51, for example.

Computing the parsimony score – The Sankoff Algorithm

$$S_i^h(x) = \min_{x_{j,h}}\{C(x_{i,h}, x_{j,h}) + S_j^h(x_{j,h})\} + \min_{x_{k,h}}\{C(x_{i,h}, x_{k,h}) + S_k^h(x_{k,h})\}$$

- Intuition: First term corresponds to the branch descending from node i to node j. This branch contributes to the length of the subtree descending from node i in two ways: first, it contributes a length along the branch connecting nodes i and j; second, it contributes a length due to the subtree descending from j, as recorded by the S function for node j.

Computing the parsimony score – The Sankoff Algorithm

$$S_i^h(x) = \min_{x_{j,h}}\{C(x_{i,h}, x_{j,h}) + S_j^h(x_{j,h})\} + \min_{x_{k,h}}\{C(x_{i,h}, x_{k,h}) + S_k^h(x_{k,h})\}$$

- Intuition: There is then a similar contribution from the other branch descending from node i, denoted by k here.

Computing the parsimony score – The Sankoff Algorithm

$$S_i^h(x) = \min_{x_{j,h}}\{C(x_{i,h}, x_{j,h}) + S_j^h(x_{j,h})\} + \min_{x_{k,h}}\{C(x_{i,h}, x_{k,h}) + S_k^h(x_{k,h})\}$$

- Intuition: Taking the minimum over all possible assignments of states to the nodes j and k will give the minimum at node i, given that it has state x.

Computing the parsimony score – The Sankoff Algorithm

- Set the S function at the external nodes:

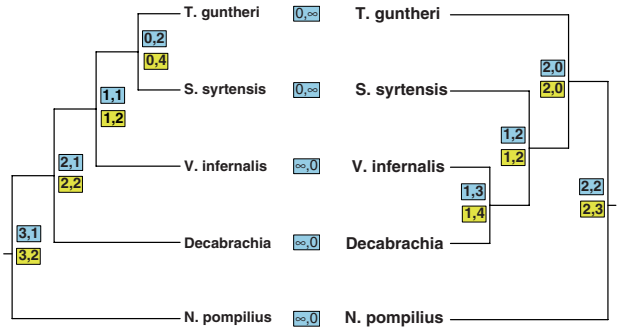
$$S_m^h(x) = \begin{cases} 0 & x_{mh} = x \\ \infty & otherwise \end{cases}$$

- Visit the nodes in a post-order traversal and compute S for each
- The minimum score is given by

$$S(\tau) = \sum_{h=1}^{N} \min_{x} S_r^h(x)$$

where r denotes the root node.

Computing the parsimony score – The Sankoff Algorithm



(a) For the first cost function, $S = 1$, while for the second cost function, $S = 2$.

(b) For both cost functions, $S = 2$.

Two sets of cost functions: Blue (upper boxes): C(0,0) = C(1,1) = 0; C(0,1) = C(1,0) = 1 Yellow (lower boxes): C(0,0) = C(1,1) = 0; C(0,1)=1; C(1,0)=2

Other important ideas in the parsimony framework

- It is sometimes desirable to obtain a most parsimonious reconstruction – an assignment of states to the internal nodes of the tree that achieves the minimum length possible on that tree
- This gives us a method to compare any set of trees. We haven't yet discussed how to find the tree (or set of trees) with minimum length for a given data matrix **X**.
- Our examples so far considered only binary data. Straightforward to extend to sequence data.

2.4 Parsimony and Consistency

Parsimony and statistical consistency

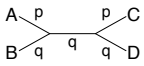
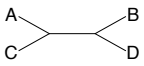
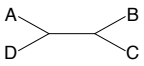
- One desirable property of an estimator is that it be consistent – as more data are added, the estimator becomes increasingly likely to obtain the true value of the quantity being estimated.

- What does consistency mean here? As sequence length increases, we become more likely to estimate the true underlying tree that generated the sequence data.

- A criticism of parsimony is that it has been shown that the criterion is not consistent under certain conditions.

Parsimony and statistical consistency

- Example data

			
0000, 1111	0	0	0
0001, 1110 0010, 1101 0100, 1011 1000, 0111	1	1	1
0011, 1100	1	2	2
0101, 1010	2	1	2
0110, 1001	2	2	1

Parsimony and statistical consistency

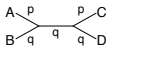
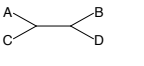
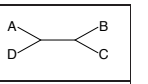
- Let p and q be the probabilities of changes along a branch

- Note that we can use these probabilities to find the probabilities of the site patterns

$$
\begin{aligned}
P_{xxyy} &= (1-p)(1-q)[q(1-q)(1-p)+q(1-q)p]+pq[(1-q)^2(1-p)+q^2p] \\
P_{xyxy} &= (1-p)q[q(1-q)p+q(1-q)(1-p)]+p(1-q)[p(1-q)^2+(1-p)q^2] \\
P_{xyyx} &= (1-p)q[(1-p)q^2+p(1-q)^2]+p(1-q)[q(1-q)p+q(1-q)(1-p)]
\end{aligned}
$$

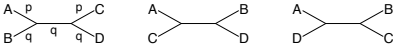
where $x, y \in \{0, 1\}$ and $x \neq y$.

Parsimony and statistical consistency

			
0000, 1111	0	0	0
0001, 1110 0010, 1101 0100, 1011 1000, 0111	1	1	1
0011, 1100	1	2	2
0101, 1010	2	1	2
0110, 1001	2	2	1

- Note that site patterns in category 1 and 2 are not informative for selecting the tree (not phylogenetically informative)
- Site patterns 3, 4, and 5 each favor a different tree

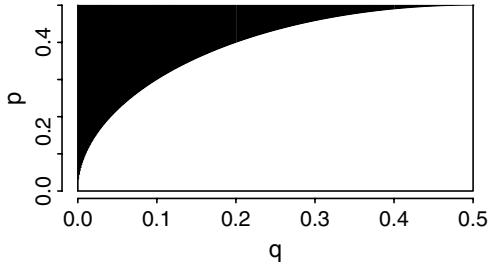
Parsimony and statistical consistency

			
0000, 1111	0	0	0
0001, 1110 0010, 1101 0100, 1011 1000, 0111	1	1	1
0011, 1100	1	2	2
0101, 1010	2	1	2
0110, 1001	2	2	1

- Suppose that the first tree listed is the true tree. Then parsimony will be statistically consistent whenever the pattern $xxyy$ occurs with higher probability than the other two.
- Using the site pattern probabilities, we can show that this occurs whenever $q(1-p) > p^2$.

Parsimony and statistical consistency

- The Felsenstein zone



- The parsimony method will be inconsistent for branch lengths in the shaded zone.
- This has also been called long branch attraction.

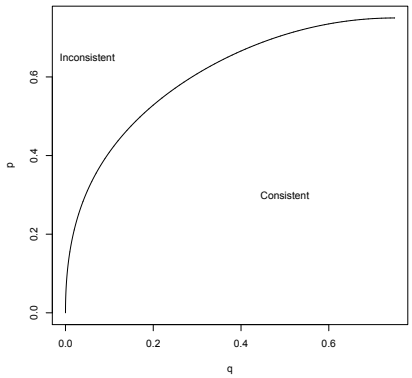
Parsimony and statistical consistency

- Can extent this to a four-state (DNA sequence) model, and find the following condition for consistency

$$p < \frac{-18q + 24q^2 + \sqrt{243q - 567q^2 + 648q^3 - 288q^4}}{9 - 24q + 32q^2}$$

Parsimony and statistical consistency

- This leads to the region below (note that the region of consistency is greater than for the two-state model)



3 Summary

Summary

- Introduction to concepts of phylogenetics: trees, data, objectives
- Introduction to our first criterion for estimation of a phylogeny
- Discussion of advantages and disadvantages